MASTER'S THESIS

In the Eye of the Beholder…

A mixed methods research on interpretation, appreciation and use of the dashboard ‘Category Analysis’ for mathematics in Cito LOVS computer program.

Hartgers, Floris

Award date:
2019

In the Eye of the Beholder…

A mixed methods research on interpretation, appreciation and use of the dashboard ‘Category Analysis’ for mathematics in Cito LOVS computer program.

In het oog van de waarnemer…

Een ‘mixed methods’ studie over interpretatie, waardering en gebruik van het dashboard ‘categorieën analyse’ voor rekenen wiskunde in het Cito LOVS computer programma

Floris Hartgers

Masterthesis Educational Sciences

Open University

Date: 13 October 2019

E-mail address: floris.hartgers@gmail.com

Name supervisor: dr. Maren Scheffel

Contents

Summary

In the Eye of the Beholder…

A mixed methods research on interpretation, appreciation and use of the dashboard 'Category Analysis' for mathematics in Cito LOVS computer program.

Floris Hartgers

Schools internationally and also in the Netherlands are stimulated to use data to enhance their education: 'data driven decision making' (DDDM). By using assessment data in a formative way, adjustment of the teaching process can be achieved. Schools for primary education in the Netherlands are obliged to use a pupil monitoring system. The tests from Cito are used in the vast majority of the schools. Most schools also use the computer program 'Cito LOVS' to analyse the gathered data from the tests. One of the possible outputs in this program, a dashboard, is the 'category analysis' for mathematics. This 'category analysis' gives the user the opportunity to examine if a pupil performs differently in certain domains of mathematics than would be expected based on the ability level the pupil achieved on the entire test. To use the information of assessments in a meaningful way, the correct interpretation of the dashboard is thus vital.

The study presented in this thesis has been executed in commission of Cito. The dashboard 'category analysis' (CA) has been in use for several years but has not been evaluated in a systematic way with users. This study aimed to research the intended interpretation of CA, the way users interpret and value CA, and to investigate if design enhancements can improve interpretation CA by its users.

The study used a mixed methods design. Cito experts were interviewed to determine the way CA was intended to be interpreted. Based on these interviews a redesign was performed. Two focus groups, each consisting of four teaching professionals in primary education, were conducted. The focus groups were shown original and redesigned dashboard examples and asked how they would interpret them. Results were used to construct a survey. The survey was completed by 278 professionals in primary education. The survey consisted of three sets of original and redesigned CA dashboards and four depicted elements of the CA dashboard. Respondents were asked multiple response questions about interpretation and multiple-choice question about knowledge of CA elements. Another part of the survey consisted of a Dutch translation of the EFLA questionnaire (Scheffel, Drachsler, Toisoul, Ternier, & Specht, 2017), to evaluate CA.

The actual interpretation of CA by users differs dramatically from the intentions as formulated by the Cito Experts. Especially the interpretation of pupil profiles that are indicated 'not prominent' cause problems. According to Cito experts these profiles should not be investigated further. Nevertheless, focus groups and over 90% of respondents in the survey indicate they would investigate these profiles

in spite of this indication. Redesign of the CA dashboard did not lead to improved interpretation by users.

This study suggests the use of CA leads to over-signalling of problems and might even lead to giving wrong or unnecessary remediation to the test taker. A radical redesign of CA without the use of graphics could be a next step to investigate. This study also supports the view that test reports and dashboards should be field tested, comparable to the testing of test/quiz/survey items, during the design phase and before distributing and implementing them in a running system.

Keywords: score report interpretation, data driven decision making (DDDM), dashboard design.

Samenvatting

In het oog van de waarnemer

Een ‘mixed methods’ studie over interpretatie, waardering en gebruik van het dashboard ‘categorieën analyse’ voor rekenen wiskunde in het Cito LOVS computer programma.

Floris Hartgers.

Scholen in binnen en buitenland worden gestimuleerd data te gebruiken om de opbrengsten te verhogen: ‘opbrengsgericht werken’. Door toetsen op een formatieve manier te gebruiken kan het onderwijs verbeterd worden. Scholen voor primair onderwijs in Nederland zijn verplicht een leerlingvolgsysteem te gebruiken. De toetsen van het Cito leerlingvolgsysteem worden door de meeste scholen in Nederland gebruikt. Het merendeel van de scholen gebruikt het computerprogramma ‘Cito LOVS’ om de toetsgegevens te analyseren. Een van de mogelijke rapporten, een dashboard, van dit programma is de ‘categorieënanalyse’ voor rekenen wiskunde. Deze ‘categorieënanalyse’ geeft de gebruiker de mogelijkheid om te onderzoeken of een leerlingen anders presteert op bepaalde rekendomeinen dan verwacht zou worden, op grond van het vaardigheidsniveau van de leerling over de gehele toets. Om de informatie uit de toets op een zinvolle manier te kunnen gebruiken, is correcte interpretatie van het dashboard essentieel.

Het onderzoek van deze thesis is uitgevoerd in opdracht van Cito. Het dashboard ‘categorieënanalyse’ (CA) wordt al jaren gebruikt, maar is nog niet systematisch geëvalueerd met gebruikers. Het doel van deze studie is te onderzoeken wat de bedoelde interpretatie van CA is, hoe gebruikers CA interpreteren en waarderen en onderzoeken of verbeteringen in het ontwerp de interpretatie van CA kan verbeteren.

Het onderzoek heeft een ‘mixed methods design’ gebruikt. Cito experts werden geïnterviewd om te bepalen hoe CA geïnterpreteerd zou moeten worden. Een herontwerp werd gebaseerd op deze interviews. Er vonden twee focusgroepen plaats met elk vier onderwijsgevenden uit het primair onderwijs. Deze focusgroepen kregen originele en aangepaste CA dashboards te zien en werden gevraagd hoe ze deze dashboards interpreteerden. De resultaten werden gebruikt om een vragenlijst te construeren. De vragenlijst werd compleet ingevuld door 278 onderwijsprofessionals uit het primair onderwijs. De vragenlijst bestond uit drie sets van originele en aangepaste CA dashboards en vier illustraties van elementen uit het CA dashboard. Respondenten kregen ‘multiple response’ vragen over interpretatie en ‘multiple choice’ vragen over kennis van elementen van het CA dashboard. Een ander deel van de vragenlijst bestond uit een Nederlandse vertaling van de EFLA vragenlijst (Scheffel, Drachsler, Toisoul, Ternier, & Specht, 2017), om CA te evalueren.

De interpretatie van het CA dashboard door gebruikers verschilt enorm van de bedoelde interpretatie zoals geformuleerd door de experts van Cito. Vooral de interpretatie van profielen die worden aangemerkt als ‘niet opvallend’ veroorzaakt problemen. Cito experts stellen dat deze profielen

niet verder onderzocht zouden moeten worden. Maar ondanks deze indicatie, geven focusgroepen en ruim 90% van de respondenten van de vragenlijst aan, dat deze profielen voor hen aanleiding geven voor verder onderzoek. Herontwerp van het CA dashboard leidde niet tot verbeterde interpretatie door de gebruikers.

Deze studie wijst erop dat het gebruik van CA mogelijk leidt tot over signalering van problemen en kan leiden tot het geven van verkeerde of onnodige remediëring van de getoetste leerlingen. Een radicaal herontwerp zonder gebruik van grafieken zou onderzocht kunnen worden. Deze studie ondersteunt ook de visie dat testrapportages en dashboards in het veld getest zouden moeten worden, vergelijkbaar met het testen van toets/test/survey-items, tijdens de ontwerpfase en voor de implementatie in een bestaand systeem.

Keywords: interpretatie score-rapporten, opbrengstgericht werken, dashboard ontwerp.

1. General Introduction

1.1 Problem and Aims of the Study

In the Netherlands, the government policy towards primary education is to maximise the learning results on main subjects like mathematics and reading. To achieve these results "opbrengstgericht werken" is propagated (Inspectie van het Onderwijs, 2010; Ledoux, Blok, Boogaard, & Krüger, 2009). The Dutch term 'opbrengstgericht werken' can be traced back to 'data-driven decision making' (Ledoux et al., 2009). Data-driven decision making (DDDM) can be defined as a cyclical process of "systematically analyzing existing data sources within the school, applying outcomes of analyses to innovate teaching, curricula, and school performance, and, implementing (e.g. genuine improvement actions) and evaluating these innovations" (Schildkamp & Kuiper, 2010a, p. 482). To generate objective data, 90% of Dutch primary schools use the tests from 'Cito leerlingvolgsysteem'. To analyse these data the majority of the schools have access to the software package 'Cito LOVS' (BTC Media Test BV.,2018; Ledoux et al., 2009).

Gathered data can be used to determine achievements (summative) or to enhance education during the learning process (formative). Especially the formative form of feedback is effective to establish better learning results (Ledoux et al., 2009). 'Cito leerlingvolgsysteem' tests for mathematics are designed to provide teachers with the level and development of numeracy of individual pupils and groups of pupils, in comparison with the average level of numeracy of their year group and to measure the development over a longer period (Hop & Engelen, 2017). This could both be interpreted as summative testing (comparing level to year group) and as formative testing (measuring development during the learning process) (Van der Kleij, 2013). The tests can also be used to research content areas in which the pupils score higher or lower than expected by their overall test-score. This analysis is done by the 'category analysis' (CA) in the computer program 'Cito LOVS' (Hop & Engelen, 2017). These outcomes could also be used in a formative way. Hattie and Brown (2007) state that school-based assessment instruments should be aligned with the school curriculum in order to give educators information with which they can improve education. The CA gives the educator a deeper view into the results with respect to different areas of the curriculum. Therefore it is expected that it can be a useful tool to improve education when used in an effective way. This study will focus on the individual CA of the tests 'Rekenen en wiskunde 3.0' for Dutch primary education.

It is important for acceptance and effectiveness of data-driven decision making that users understand the data (Schildkamp & Kuiper, 2010a). Dutch inspectorate of education and Cito employees observe that educators often lack the proper 'data-literacy' to interpret the test results correctly (Ledoux et al., 2009). To make data accessible for all stakeholders the use of transparent graphics can help stakeholders to interpret data, without being assessment literate in a classical sense (Hattie & Brown, 2007). The 'LOVS tests' provide teachers with numerical information and graphical

representations. Van der Kleij (2013) found that educators often misinterpret results from several graphical reports of 'Cito LOVS.' She studied and enhanced different score reports of 'Cito LOVS' (Van der Kleij, 2013). She did not, however, investigate or improve the 'category analysis,' which gives more detailed information about the skills of pupils with respect to different mathematical content areas.

This study was performed in commission of Cito. The aim of this study was to investigate how the category analysis instrument of 'Cito LOVS' for the test 'Rekenen-Wiskunde 3.0' is interpreted, appreciated and used by educators and to give possible alternatives to provide teachers with graphics they can interpret and use in a meaningful way.

1.2 Theoretical Framework

1.2.1 Data feedback

Data Driven Decision Making (DDDM) is the systematic collection, registration, analysis and interpretation of data to inform decision making in an educational setting (Ledoux et al., 2009; Mandinach, 2012; Van der Kleij & Eggen, 2013). This interpretation of DDDM broadly overlaps with the term Learning Analytics: "Learning analytics (LA) is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs."[1] This term has been used more and more during the last years and takes into account the growing amount of data being gathered in education (Picciano, 2012; Tempelaar, Heck, Cuypers, van der Kooij, & van de Vrie, 2013).

According to Davenport and Prusak (1998, p. 2), data is "a set of discrete, and objective facts about events", with little meaning in itself. When meaning is added to data it becomes information which can be used. Data that has been gathered must be translated into understandable information for educators, in order to be transformed into actions that improve instruction (Mandinach, 2012; Schildkamp & Kuiper, 2010b).

A frequently used way of gathering data is through assessment (Van der Kleij, 2013). Schools can use summative and formative assessment. Summative assessment is used to evaluate learning progress in hindsight. When assessment is used to enhance learning during the learning process, this can be seen as formative assessment (Ledoux et al., 2009). This strict distinction is nuanced by Bennett (2011), who states that formative assessment has an emphasis on 'assessment *for* learning', but also measures 'assessment *of* learning', as opposed to summative assessment with an emphasis on 'assessment *of* learning'. There is reason to believe formative assessment can contribute to the improvement of student performance (Cavalluzzo et al., 2014; J. Meijer, Ledoux, & Elshof, 2011), but the effects may vary due to different forms and implementations of formative assessment (Bennett,

[1] 1st International Conference on Learning Analytics & Knowledge: https://tekri.athabascau.ca/analytics/

2011). A central aspect of effective formative assessment seems to be the inference taken from the observed errors. An error can be made by students due to a slip or a misconception, which require different actions to be taken by the instructor. To discover what caused the error, an extra form of assessment is needed, for example asking the student to explain why a mistake was made. With this information, the instructor can discover if a simple feedback remark or extra instruction is needed (Bennett, 2011).

Important tools for educators to translate data into meaningful information are pupil-monitoring systems, which can be used for educational decision-making (Van der Kleij & Eggen, 2013). According to Faria et al. (2012) these tools can be meaningful for improved student achievement when used for three general purposes: better understanding of and response to academic needs of individual students; better understanding of instructional capacities of individual teachers; support and facilitation of conversations among teachers and instructional leaders to improve instruction. Hattie, Brown, and Keegan (2003) describe assessment as powerful when it provides teachers with information about the target of learning, the actual progress in relation to the target and directions related to future teaching.

To be able to use the gathered data, educators have to develop 'data literacy'. They have to understand data in a way they can transform it into relevant information which can be used effectively to inform decisions (Mandinach & Gummer, 2013). However, several studies report the lack of data literacy of school-staff (e.g. Mandinach, 2012; Mandinach & Gummer, 2013; Schildkamp, Karbautzki, & Vanhoof, 2014; Vanhoof, Verhaeghe, Verhaeghe, Valcke, & Van Petegem, 2011), and there is reason to believe Dutch educators also lack a high degree of data literacy (Schildkamp et al., 2014; Van der Kleij, 2013). Many educators have trouble interpreting terminology or graphical displays used in reports. Especially measurement error is a difficult area to understand (Van der Kleij & Eggen, 2013; Zapata-Rivera, Zwick, & Vezzu, 2016). Although many studies call for training of educational staff (e.g. Mandinach & Gummer, 2013; Schildkamp et al., 2014), this is beyond the scope of this thesis. Given this situation, it is important the data are presented in a form that enables educators to actually use it in a meaningful way (J. Hattie, 2009; Hopster-den Otter, Wools, Eggen, & Veldkamp, 2017; O'Leary, Hattie, & Griffin, 2017).

1.2.2 Dashboards

Educators are able to make use of test scores when these scores are reported in a useful, accessible and understandable way (Hambleton & Zenisky, 2013; O'Leary et al., 2017; Van der Kleij, Eggen, & Engelen, 2014). The test results have to be presented in a context that enables stakeholders to give practical meaning to them (Hambleton & Zenisky, 2013). The correct interpretation of the presented results by the readers is of the utmost importance and affects the validity of usage of the test report (Gotch & Roduta Roberts, 2018; Hattie, 2009; O'Leary et al., 2017). Also, correct interpretation of test

results is a necessary precondition to use assessment results to enhance education (Van der Kleij et al., 2014). In spite of this, too little attention has been paid to the fitting of test reports in relation to data and the actions audiences take based on this (Hopster-den Otter et al., 2017). Test reports should be field tested, comparable to the testing of test/quiz/survey items (Ryan, 2006). Ryan (2006) recommends the use of focus groups to test reports in relation to the users.

A way of reporting test results is to make use of a learning dashboard. "A learning dashboard is a single display that aggregates different indicators about learner(s), learning process(es) and/or learning context(s) into one or multiple visualizations" (Schwendimann et al., 2017, p. 37). To use dashboards effectively, Verbert, Duval, Klerkx, Govaerts and Santos (2013) developed a four-stage learning analytics process model: awareness (data), reflection (questions), sense making (answers) and impact (behaviour change). When graphics are used, it enables users to access the deeper structure of reported data without having to engage in complicated decoding of numbers or text. Graphics can lead the reader straight to the meaning, while numbers are merely indicators (Hattie & Brown, 2007). To effectively use the presented data, graphics help the reader to make sense of the data and give them impact.

To improve the formative use of test scores, subscores that indicate strengths and weaknesses of students can be helpful (Goodman & Hambleton, 2004; Hopster-den Otter et al., 2017; Monaghan, 2006; Van der Kleij et al., 2014). However, the use of subscores raises some concerns about the reliability of the reported results (Hambleton & Zenisky, 2013). Most subdomains contain limited amounts of items (Goodman & Hambleton, 2004) and scores of subdomains often are redundant with the total test score (Monaghan, 2006). Using untrustworthy information from subscores may result in misinterpretation and giving wrong remediation to the test taker (Hambleton & Zenisky, 2013; Monaghan, 2006). Also, there is the risk of a completely different profile on a retest when subscores are presented with low reliability. This may result in confused users, or even in ‘blaming the test’ for being inconsistent (Twing, 2008). Users have problems interpreting standard errors and reliability of reported subscores. They should be given clear guidance on how to interpret and use the test results (Goodman & Hambleton, 2004).

As stated before, graphics can help users of dashboards to understand the deeper structure of complex data (Hattie & Brown, 2007). However, graphics give highly salient and persuasive information and may lead to overuse of numbers. Alternatively, tables or text can be used instead of graphics, because they are relatively ‘neutral’ to interpretation (Zapata-Rivera & Zwick, 2011).

People have difficulties interpreting proportions in graphics (Shah, Freedman, & Vekiri, 2005). Also, graphics can be constructed in a way that makes them difficult to interpret. Kosslyn (2006) formulates eight psychological principles to communicate a message to a specific audience by using graphics: relevance, appropriate knowledge, salience, discriminability, perceptual organization, compatibility, informative changes and capacity limitations.

1.2.3 Cito LOVS category analysis

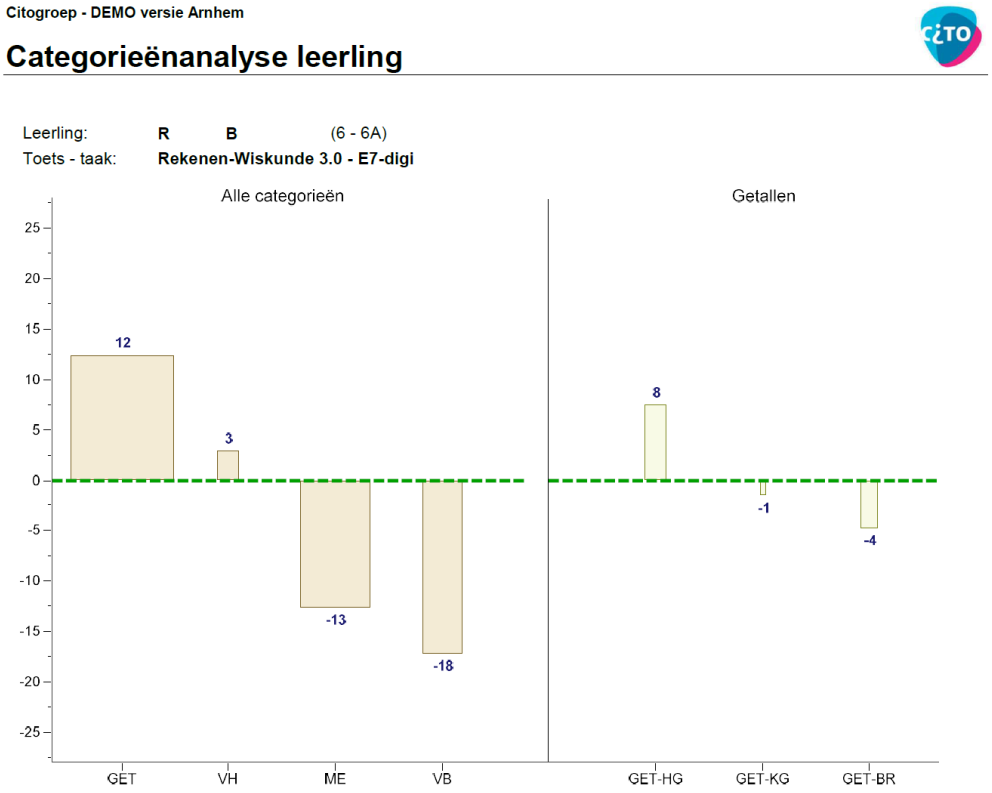
In the Netherlands 'Cito-LVS' is the most common pupil-monitoring system in primary education. It provides biannual tests for mathematics, reading and writing. 'Cito-LVS' tests are used by approximately 90% of the schools; the majority of the schools also use the computer program 'Cito-LOVS' to help analyse test results (BTC Media Test BV.,2018; Ledoux et al., 2009). This computer program can be used on school level, group level and individual level. The assessment system has a statistically strong foundation (Meijer et al., 2011) and can be used in a summative and formative way (Van der Kleij, 2013). The tests and computer program are capable of producing a lot of different analyses and outcomes. The test results are being used to monitor the development of pupils. Results are hardly ever used for feedback or as a basis to make decisions (Hopster-den Otter, Wools, Eggen, & Veldkamp, 2017; Meijer et al., 2011; Schildkamp & Kuiper, 2010).

School leaders, internal support teachers and teachers using the computer system have a lot of difficulties interpreting the results in a correct way (Meijer et al., 2011; Van der Kleij & Eggen, 2013). Teachers have difficulties interpreting the dashboards on individual and group level. Internal support teachers seem to be better able to interpret the dashboards than teachers and school-leaders (Van der Kleij, 2013). Experience using the Cito LOVS computer program, did not make a difference in interpreting the reports (Van der Kleij & Eggen, 2013). The effects of training are not clear yet. Van der Kleij & Eggen (2013) did not find a significant effect for training, but they did not investigate what kind of training was done by different respondents. Staman, Visscher & Luyten (2014) investigated the effect of a training program of seven sessions and showed a positive effect of training on skills of data-driven decision making and interpreting score reports. All three user-groups struggled with the concept of confidence interval. They did not find information related to confidence intervals useful when they were visualised in a score report. (Van der Kleij & Eggen, 2013).

One of the dashboards available in the 'Cito LOVS' computer program is the CA for mathematics (see figure 1). This CA analyses the Cito LVS test for mathematics. The Cito LVS test for mathematics is based on the assumption that the ability for math can be seen as a unidimensional continuum (Janssen, Verhelst, Engelen and Scheltens, 2010). With CA, it is possible to see if students, given their current level, performed in a balanced way on the different subdomains (Cito, 2018). It does not provide the reader with subscores per domain, but produces a profile of domains that is compared to the expected profile, based on the total test score. This profile of different subdomains is distinguished within the unidimensional test scale. The subdomains are distinguished on didactical grounds (Verhelst, 2007) and differ from test to test (Cito, 2018). This instrument differs from classical error analyses in that the latter counts amounts of right responses per category and does not take the general level of the student into account (Janssen & Hickendorff, 2008). The second half of the bar graph provides the reader with a profile based on one of the subdomains, which is divided in a

profile of further subdomains (see figure 1). For example the dimension ‘getallen’ [numbers] can be divided into ‘hele getallen’ [whole numbers], ‘kommagetallen’ [decimals] and ‘breuken’ [fractions].

To determine the scores on subdomains each test-item is given a different ‘weight’ due to the difficulty level and discrimination ability of the item. The ‘item-weight’ is determined based on the ‘Item Response Theory’ (IRT) (Janssen & Hickendorff, 2008). The scores on the subdomains are compared to the expected scores of students with the same overall test score. The differences between the observed profile and the expected profile are quantified in a chi-square distance. Each domain contributes to the total chi-square distance. The contribution of each domain to the total chi-square distance differs due to the number and weight of the items per domain. The difference between an observed score on a domain and the expected score on a domain are presented in the dashboard as a percentage ‘score deviation’ in a table and a bar graph (Janssen & Hickendorff, 2008).



Categorie label	Categorie	Aantal opgaven	% Score geobserveerd	% Score verwacht	% Score afwijking
GET	Getallen	40	85	73	+12
VH	Verhoudingen	16	79	76	+3
ME	Meten	28	55	68	-13
VB	Verbanden	12	55	73	-18
GET-HG	Getallen - Hele getallen	13	93	85	+8
GET-KG	Getallen - Kommagetallen	14	84	85	-1
GET-BR	Getallen - Breuken	13	80	84	-4

Figure 1. Demonstration of CA, retrieved from Cito BV.

If the likelihood of getting a certain chi-square distance on the entire profile is less than 10%, this profile is labelled as ‘prominent’; when the chance of observing a chi-square distance is less than 5%, the profile is labelled as ‘very prominent’. These indications are a sign the profile should be investigated further. This label is indicated by the horizontal graphs under the bar graph. These graphs also indicate if a profile is nearly ‘prominent’ by filling the horizontal bar (Cito, 2018). The expected profiles are theoretically constructed, based on the calibration of the individual test items using IRT (Janssen & Hickendorff, 2008). Overall test-results within the highest and lowest 10% are not used for this procedure. The Cito LOVS user manual points out that profiles that are indicated with ‘not prominent’, but are almost ‘prominent’ –which is indicated by the horizontal bar– are worth analysing as well (Cito, 2018).

The chi-square distances model was tested with the outcomes of the ’Cito-eindtoets’. On grounds of the model it was expected that 10% of the test would be classified as significantly different; it turned out that 14% of the profiles were marked as significantly different. This exceedance can be acceptable, provided that it is adequately communicated to the educational field (Verhelst, 2007).

The most difficult task is formulating advice based on the profile. Striking patterns can be caused by coincidence and do not necessarily have to point out a problem (Verhelst, 2007). Interpretation should be done with great care. Striking patterns can be caused by several reasons, which are not explained by the CA and should be further analysed (Cito, 2018; Janssen & Hickendorff, 2008). A striking pattern can be discovered for an individual pupil, but if a large amount of the pupils in a class or school perform different than expected, this can also be a reason to investigate the reasons for the unexpected outcomes (Janssen & Hickendorff, 2008).

1.3 Research Questions

The overarching research question of this study is: “Is the ‘Cito category-analysis’ a tool that gives meaningful information to teachers and internal support teachers in Dutch primary education?” This question was assessed by consulting experts from Cito and after this, organising focus groups of teachers and internal support teachers. In the focus groups the present use of the dashboard was evaluated. Furthermore, the dashboard was refined based on the outcomes of the interviews and focus groups. Findings of the focus groups were tested via a survey about using the CA with educators in Primary Education in the Netherlands.

The following sub-questions guided the research conducted for this thesis:

1. What is the intended interpretation of the ‘Cito category-analysis’ dashboard?
2. Which enhancements or alternatives can be designed, based on expert opinion, users’ views and literature, to help users interpret the ‘Cito category-analysis’ dashboard in an appropriate way?

3. How is the ‘Cito category-analysis’ evaluated by experienced and intermediate users?
4. Is the actual interpretation of the ‘Cito category-analysis’ dashboard by teachers and internal support teachers in accordance with the intended interpretation?
5. Is the interpretation of the alternative dashboard by teachers and internal support teachers more in accordance with the intended interpretation?
6. Is there a difference in interpretation of the ‘Cito Category Analysis’ dashboard and the alternative dashboard between experienced users, intermediate users and new users, or between teachers, internal support teachers and school leaders?
7. What guidelines can be deduced for a future redesign of the ‘Cito Category Analysis’ in particular and graphics in pupil monitoring systems for educators in general?

In figure 2 the relations between the questions are graphically displayed.

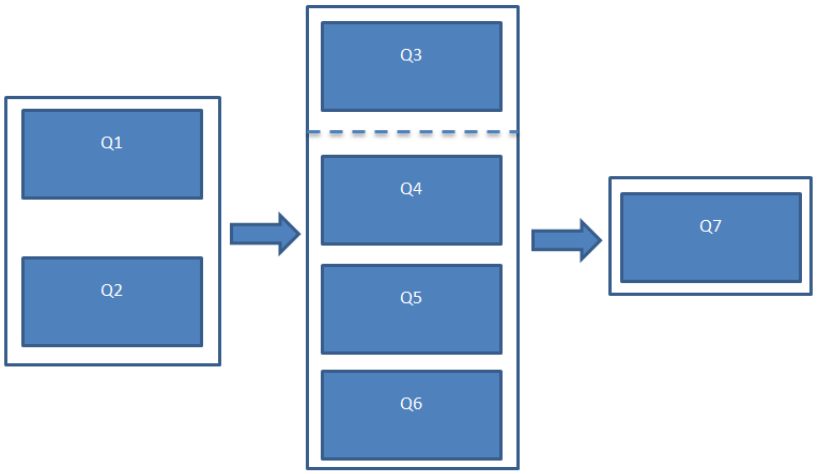


Figure 2. Research model

2. Method

2.1 Design

Commissioned by Cito, this study investigated how the “Cito LOVS Category Analysis for mathematics” is interpreted, appreciated and used by teachers and internal support teachers in Dutch primary education. By using a mixed methods design it was possible to gain a deeper understanding of the subject and triangulate findings (Creswell, 2014; Johnson & Onwuegbuzie, 2004). Because this subject has not been researched before an ‘exploratory sequential design,’ as shown in figure 3, made it possible to identify themes by qualitative data collection, which could be tested in a quantitative way (Creswell, 2014; Meijer, Verloop, & Beijaard, 2001).

By interviewing experts and focus groups of users, different themes were identified. The use of focus groups to evaluate assessment reports is recommended by Hattie (2009), Ryan (2006) and Van

der Kleij (2013). By using focus groups, a lot of information from several users can be gathered in a relative short time (Creswell, 2014).

With the themes ascertained by the expert and focus group interviews, a part of a survey was constructed. Another part of the survey contains a Dutch translation of the questions of the EFLA questionnaire, which –in English– is validated for evaluating learning analytics tools (Scheffel, Drachsler, Toisoul, Ternier, & Specht, 2017). With the use of this cross-sectional survey among a bigger group of users of 'Cito LOVS', findings of the focus group meetings were triangulated and validated (Creswell, 2014).

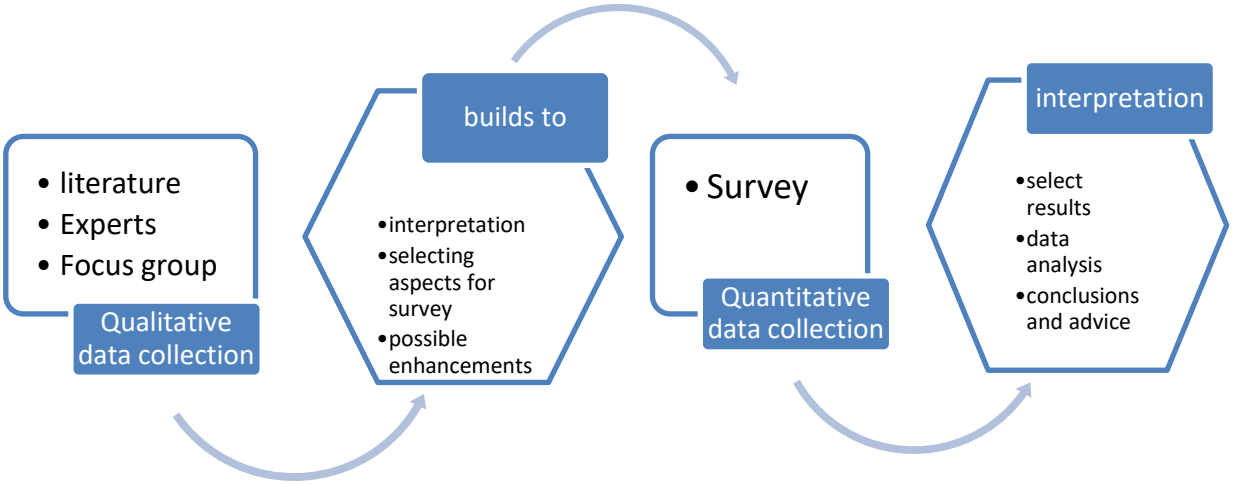


Figure 3. Model of the exploratory sequential design

2.2 Participants

2.2.1 Cito experts

Three experts from Cito have been consulted about the intended use of the 'Category Analysis'. One expert is an advisor for primary education, one is an information manager and one is a statistics expert. All three have had long experience with this particular dashboard. For the design of the survey a test-expert and a database marketeer cooperated.

2.2.2 Focus groups

Two focus groups of four people were formed to investigate the use and interpretation of the CA dashboard. The focus groups each consisted of colleagues working in a school for regular primary education in large cities in the middle and West of The Netherlands. One school was populated by children with a low social and economic status background, one school had a mixed population. Both schools of the focus groups worked 'data-driven', using 'Cito LVS-tests'. All participants were experienced users of CA.

The first focus group contained three females and one male; they have worked in primary education between 9 and 25 years. One member was internal support teacher, three members taught classes; one of the teachers also was ICT-coordinator of the school. The second focus group was also formed by three females and one male. Their experience in primary education varied between 7 and 21 years. Two participants of this focus group were internal support teachers, two persons taught classes. One of the teachers and one of the internal support teachers coordinate the math-education within the school.

2.2.3 Survey

In the Netherlands there are 6626 schools for primary and special primary education with an average of 9 to 10 different classes (schoolyear ’16-’17) (Inspectie van het Onderwijs, 2018). Of the teachers, 87% are female (Beiro & Ramaekers, 2016). Of the teaching staff 32% are aged under 35 years, 24% are aged between 35 and 45 years, 21% are aged between 45 and 55 years and 22% are aged 55 years and older (Centraal Bureau voor de Statistiek, 2019a). Of the schools, 10% are located in the North of the Netherlands, 21% in the East, 48% in the West and 21% in the South of the country (Centraal Bureau voor de Statistiek, 2019b). Teachers in The Netherlands report a high workload (Inspectie van het Onderwijs, 2018), which can have a negative impact on participating in this research. Over 50% of the schools use the ‘Cito LOVS’ computer program and have access to the ‘category analysis dashboard’ (BTC Media Test BV, 2018).

The survey was first distributed to 4,208 unique email contacts of ‘Cito Portal administrators’ within schools; their other functions within their institutions are not known to Cito. Teachers and internal support teachers were invited to participate. After one week 101 respondents had completed the survey, which is a response rate of 2.4%. It was decided to send a reminder to the contacts that had not reacted. Also, a new invitation was sent to another database of known Cito contacts. This database was checked on convergence with the first used database and contacts that were registered in both databases were deleted. After this procedure the second database contained 2,139 unique email contacts, of which 1,554 are subscribed to the Cito newsletter, 460 are contacts who indicated they wanted to react on developments of new Cito products and 125 are members of the ‘Cito community’. After another week, a total of 278 surveys had been completed by unique respondents, which is an overall response rate of 4.4%.

Of the respondents, 86% indicated to be female (total N=272, missing=6), 9% worked for a school located in the North, 21% in the East, 48% in the West and 21% in the South of the Netherlands (total N=278). This is comparable to the entire population. Of the respondents 13% indicated they were aged under 35 years old (total N=277, missing=1), which is much lower than the population. The group of respondents was formed by 59 teachers (including teachers with an additional task, such as ICT or math coordinator), 192 internal support teachers (including internal

support teachers that indicated they were also teacher), 22 managers and 5 respondents who indicated they had other functions such as educational advisor (total N=278).

Within the survey half of the respondents was randomly assigned to read a warning message about chance (see materials section). Of the respondents who completed the survey 45% (n=125) was shown the message and 55% (n=153) did not see this message. Of the respondents that completed the survey 82% (n=228) indicated they were familiar with CA; of these respondents 77% (n=176) indicated they used CA at least once a year. These 176 CA users were presented with the EFLA questionnaire.

As the databases consists of older respondents than average with a large group of specialised educational professionals, there is a possibility of bias in the response group. It is also possible that people with more knowledge about Cito tests and more interest in Cito tests or using data, are overrepresented. Also, there is a chance that people that are more interested, or have a more positive attitude towards testing, will be more inclined to respond. Within this study it was not possible to prevent this**.**

2.3 Materials

2.3.1 Cito experts

The expert meetings were conducted as semi-structured interviews. Several demonstration displays of the CA dashboard were printed. Questions in this interview were for example: “Who is the target group for the CA for Cito LVS Mathematics?”, “How should this dashboard be interpreted?” The interviews were transcribed by using ‘Listen N Write free; 1.30.03’. The transcriptions were analysed using ‘RQDA’ (Copyright (c) 2008-2009, Ronggui Huang), an open source tool for qualitative data analysis in ‘R’. To determine the aims of the CA dashboard, the Cito LOVS manual (Cito, 2018) and descriptions of the category analysis (Janssen & Hickendorff, 2008; Verhelst, 2007) were used in addition to the information from the experts.

2.3.2 Focus groups

For the focus groups five displays of the CA dashboard were produced, also five alternative variations of these dashboards were produced (see Appendix A). These original displays were made for demonstration purposes by Cito and cannot be linked to real persons. Examples of ‘not prominent’, ‘prominent’ and ‘very prominent’ profiles in different combinations were printed, one subdomain had the message it could not be calculated due to a high score. The alternative dashboards were produced based on interviews with experts and eight principles for graph design (Kosslyn, 2006). In the focus groups, all displays were printed and presented on laminated A4 sheets. Between the focus groups, slight changes were made in the alternative dashboards, based on the first focus group.

For the focus groups a protocol was made (see appendix B). In this protocol questions about the participants were formulated: “Can you introduce yourself by telling your name, your function and

how long you are working in primary education?" Questions about experience: "How long have you been using the category analysis for mathematics?" Questions about the displays, such as: "What do you see?" and "What would you conclude from this report?" Also questions about possible enhancements: "What could be made more clear in the report?" A video camera was used to record the focus groups. These interviews also were transcribed using 'Listen N write free; 1.30.03' and analysed using 'RQDA' (Copyright (c) 2008-2009, Ronggui Huang).

A consent form with information about the aims and relevance of the research, gathering and handling of the data and the video material and the transcriptions was made.

2.3.3 Survey

An online survey was constructed. The survey consisted of demographic questions, questions about experience in education and with CA, the EFLA questionnaire, and questions about interpretation, use and knowledge about displays of the dashboards (see Appendix C). The first questions contained items about experience with 'Cito LOVS' and the 'category analysis', to determine the 'routing' through the questionnaire. Respondents that indicated they did not use 'Cito LVS' tests, were presented only background questions. Respondents who knew and used CA at least once a year were presented the whole survey. Respondents who indicated they used CA incidentally were not presented with the EFLA questionnaire, but they were presented with the questions about the dashboard displays. The end of the survey consisted of background questions about gender, age, location of the school and experience in education.

A second section of the survey was the teacher version of the EFLA questionnaire that consists of eight questions in which educators can evaluate a learning analytics tool on three dimensions: data, awareness & reflection and impact. The answers can be given on a 10 level Likert scale indicating "strongly disagree" on one side of the scale, opposing "strongly agree" on the other side of the scale. The questionnaire has been validated in English (Scheffel et al., 2017). For this survey questions were translated into Dutch in coordination with Maren Scheffel, the developer of the EFLA questionnaire. The EFLA results were analysed using an interactive spreadsheet, constructed by the developers of the EFLA questionnaire. The scales of the Dutch translations, answered by 177 respondents, showed high reliability scores: Data (N=2), Cronbach's $\alpha = .81$, Awareness & Reflection (N=4), Cronbach's $\alpha = .86$ and Impact (N=2), Cronbach's $\alpha = .81$.

The third part of the questionnaire contained 10 questions about interpretation, intended use and knowledge about the CA dashboard (Cronbach's $\alpha = .63$). Three displays of the original dashboard and three displays of derived alternatives based on the same data were shown (see Appendix C for all dashboards). The six different displays were shown to the respondents in a random sequence. One set of displays gave two not-prominent profiles (dashboards 1 and 1A), one set of displays showed a very prominent profile with a the remark that a sub-profile could not be made due to a high score (dashboards 2 and 2A), and one set of displays showed a not-prominent profile with a prominent sub-

profile (dashboards 3 and 3A). The respondents were asked to select which of the presented dimensions they would like to investigate further, given the presumption the display would be of one of their pupils. It was possible to choose more than one category, it was also possible to choose not to investigate any of the dimensions. To score correct, i.e. in accordance to how Cito intended the display to be interpreted, in case of a 'not prominent' profile or a 'not prominent' part of a profile, the answer 'no further action' should be chosen. For the '(very) prominent' profiles (or part of a profile) a large negative deviation should be chosen for further investigation, to gain a point. If large positive deviations are chosen as well, the answer is still indicated as correct. In one case the sub category: 'optellen en aftrekken' displayed is 'prominent', but the category in itself is part of a 'not prominent' profile; in this case the selection of the category 'optellen en aftrekken' did not influence scoring, because it could be reasoned it is worth investigating the whole category.

Next to the six displays that had to be interpreted, four identical original display of the CA were shown with multiple choice knowledge questions about the meaning of certain aspects of the display: the green line (indicating the expected profile), the height of the bars, the surface of the bars and the meaning of the graph 'signal'. These multiple response and multiple choice were given a right or wrong norm, in total ten points maximum could be gained. As a good interpretation of the display is vital for further action, a test expert of Cito set a norm for the interpretation and knowledge questions of 90% correct answers.

Before showing the questions of this part of the survey, half of the respondents were shown a message about the CA:

> "Let op! De Categorieënanalyse geeft het verschil tussen de behaalde scores per categorie en de verwachte scores op grond van het behaalde vaardigheidsniveau weer. Er is altijd een afwijking te zien, mogelijk veroorzaakt door toeval. Alleen bij een significant verschil tussen de behaalde scores en de verwachte scores binnen het profiel zal de aanduiding 'opvallend' of 'zeer opvallend' te zien zijn." [Attention! The Category Analysis indicates the difference between achieved scores per category and expected scores per category on ground of the achieved 'skill-score'. There will always be a deviation, possibly caused by chance. Solely in case of a significant difference between achieved scores and expected scores within a profile, the indication 'prominent', or 'very prominent' will be shown.]

The survey was constructed in coordination with Cito-employees in an online survey system: 'Survey Monkey', which is contracted by Cito. The data for Cito gathered by 'Survey Monkey' is stored on servers within the European Union and can only be reached by Cito employees. An invitation email with information about the goals and relevance of the research and an invitation link to the survey was made. On the first page of the survey information about the research and the

handling of the data was given, a mandatory consent box was built in the survey. The data from the survey were analysed using the 'IBM SPSS 24' computer program. An a-priory power analysis was performed, using G*Power Version 3.1.9.2.

2.4 Procedure

2.4.1 Cito experts

The meetings with Cito experts took place in February 2019, in the Cito office in Arnhem. Two semi-structured interviews with two and one expert were held. The first meeting was with a project-leader mathematics and a data-analyst. The second meeting was with a customer supporter. The experts were informed about the aims of the research and signed a declaration of informed consent for their cooperation. The semi-structured interviews were audio-recorded and transcribed afterwards. A report of the interview was shared with the experts and slight changes were made.

2.4.2 Alternative displays

Based on the information about the intended use of CA and graph design principles formulated by Kosslyn (2006), the original display of CA was redesigned as shown in figure 4. The general aim of the redesign was focused on a more clear distinction between 'not prominent', 'prominent' and 'very prominent' profiles and to show only the information that gives meaningful information to the reader.

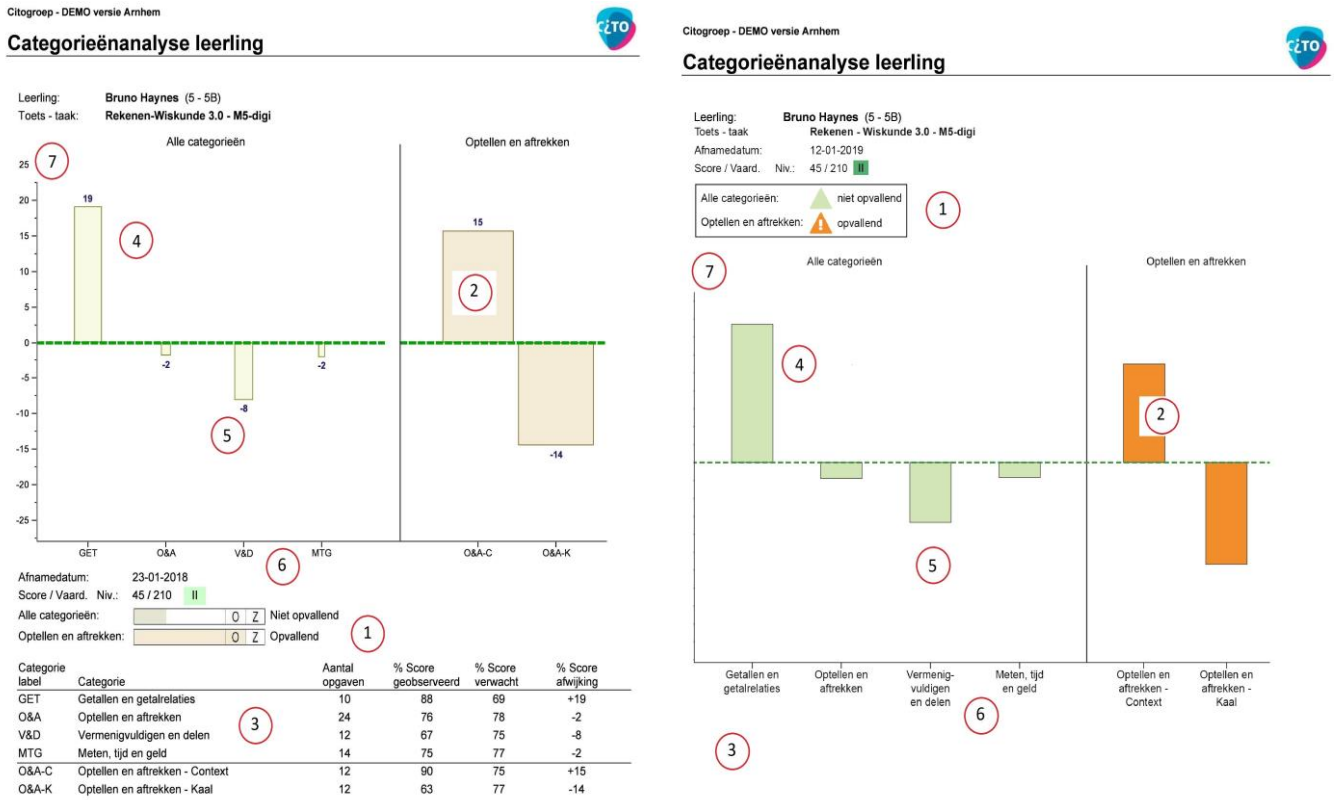


Figure 4. Changes between the original (left) and alternative (right) display of the dashboard

The ‘Principle of Relevance’, the ‘Principle of Salience’ and the ‘Principle of Discriminability’ led to a more prominent way of showing ‘not-prominent’, ‘prominent’ or ‘very prominent’, because this is the first and most important sign the reader must see. In the display, this label was made bigger (1), moved up (1) and also colour changes were made to the bars (2). The ‘Principle of Appropriate knowledge’ that states the knowledge of the user should be taken into account, led to removing the table (3) with very specific information. The ‘Principle of Perceptual Organisation’ led to elimination of the changing surfaces of the bars (4). The ‘Principle of Compatibility’ states that graphical displays should be compatible with the meaning it communicates. This led to colour changes of the bar: green for ‘not prominent’, orange for ‘prominent’ and red for ‘very prominent’ (2). The ‘Principle of Informative Changes’ that states the changing information in a display should be informative, led to removing elements that are not important to interpret the display: percentages of deviation (5). The ‘Principle of Capacity Limitations’ led to writing the whole names of the domains, instead of using abbreviations which can be found in the table (6). This last principle also supports the elimination of the table with contents (3). A last change that could be argued on grounds of the ‘Principle of Appropriate Knowledge’ and the ‘Principle of Relevance’ and the ‘Principle of Capacity’ is the leaving out of a scale on the Y-axis (7), because this scale is hard to interpret and not really informative to the user.

2.4.3 Focus groups

In March 2019 two contacts were established to form focus groups. One contact was established through an acquainted school-leader and one contact person reacted on an appeal, placed on a Facebook forum for students of the Open University. The contact persons each formed a focus group of four colleagues, including themselves. In April and May two semi-structured focus group interviews took place. The interviews were taken at the schools of the focus-group participants. The duration of the interviews was about an hour. The researcher fulfilled the role of moderator during the interviews. First, the aims of the research were explained and participants were informed, verbally and by a letter, about the way data would be handled. Participants gave a written consent for using their data. After this, the participants were asked to introduce themselves and to explain their function, experience in education and with the CA. Then five original and five alternative displays of CA were shown one by one to the focus group. The group discussed the way they would interpret the output and what they would do next if this was a real display of the results of one of their pupils. Each display was discussed for about five minutes. At the end of the interview the researcher explained some details about the intention and construction of the CA and participants were able to react on that. The whole meeting took about an hour for each focus group. Participants of the focus groups were rewarded with a voucher of €25,- provided by Cito. Both participant groups were sent a report of the meeting and were able to respond to it. They did not propose any alterations.

The focus groups were videotaped and transcribed afterwards. Gathered material was handled with great care, and stored in a locked safe, when it was not being processed. The coded material did not contain any personal information, names were replaced by a code.

2.4.4 Survey

With the outcomes of the focus group interviews a questionnaire was constructed in June 2019. The questionnaire was discussed with Cito-experts and changes were made. After the online construction of the questionnaire, it was tested by the researcher on three teachers, which resulted in small changes.

In June 2019 the invitations to the online survey were distributed by a Cito marketeer. After one week a reminder was sent and a new group of contacts was invited to participate. This second database was checked on convergence with the first database by the Cito marketeer. After another week the survey was closed and results were analysed.

2.5 Data-Analysis

2.5.1 Experts

First the interviews were transcribed. Then they were coded, using 'N-vivo' codes, using the terms and phrases of the participants (Creswell, 2014). These codes were grouped and organised into categories. Also, striking statements were marked. After this phase a core category was chosen: 'retrieving information from category analysis'. Around this core category a model was formed (figure 5), in the fashion of Grounded Theory Design (Creswell, 2014). An abstract of the interviews and the model were sent to the experts to give them the opportunity to react (Creswell, 2014). This led to some small alterations.

2.5.2 focus groups

The focus groups were transcribed and coded with 'N-vivo' codes; these codes were collapsed into categories. Striking quotes were marked. The codes and categories were plotted to get a better view on interactions between categories. This resulted in a written abstract of the focus groups with use of striking quotes. These abstracts were sent to the contact persons of the focus groups to give them the opportunity to react (Creswell, 2014), which did not lead to alterations. The most important findings were linked to the categories of the constructed model and are presented in table 1.

2.5.3 survey

The dataset was inspected. All incomplete surveys were removed, only a few surveys (n=6) that had missing values for the background questions about sex, age or region, were not removed out of the dataset. A descriptive analysis of the survey data was conducted. Respondents were described and different groups were distinguished, such as function and experience in education in general and with the CA in particular. Background questions were compared to the entire population.

The EFLA questionnaire was analysed. This was done by looking at the scores per question and calculating scale-scores. Also, reliability was checked by calculating Cronbach's α.

Answers to the multiple response and multiple choice answers were recoded into correct or incorrect. Correct answers were rewarded one point, which resulted in total scores varying from zero to ten. Subscores for the original display (n=3), the alternative display (n=3) and the knowledge questions (n=4) were calculated. Cronbach's α was calculated for these ten questions. Also each question was scored and a percentage of correct given answers of the total or subgroup was calculated. Each question and the total score was compared to the norm of 90% correct, which was set by a Cito expert.

The scores for the three original and alternative displays and each set of displays were compared using a paired-sample t-test. Scores of respondents who were shown the warning message were compared to the scores of the respondents that did not see this message using an independent t-test. Scores of four different experience groups were compared using ANOVA; for this test respondents that indicated they did not know CA and did not use CA were combined. Scores of three function groups: teacher, internal support coach and managers, were compared using ANOVA; other functions (n=5) were left out of this comparison. When confidence intervals are given, or tests are performed, a 95% significance level was used, as is common in modern research (Field, 2013).

3. Results

3.1 Cito Experts

In both meetings, the need to investigate the way users interpret CA was acknowledged by the experts. The presumption that the ability for math can be seen as a unidimensional continuum is the starting point of the analysis. Key element of the model, according to the consulted Cito experts, is the distinction between displays that are indicated as 'not prominent' or indicated as '(very) prominent'. The reader should always look at this indication first and only 'prominent' or 'very prominent' graphs should be interpreted and possibly lead to follow up actions.

The profiles indicated '(very) prominent' should be interpreted with care and lead to further investigation before action is taken, a diagnostic conversation is often recommended. The interviews resulted in a model for design and interpretation of CA (figure 5).

3.2 Focus Groups

The focus groups showed many different problems interpreting the CA dashboard in its original form (table 1). Despite this, the participants indicated they valued CA and used it often, even though they had to enter data manually into the Cito LOVS computer program. The participants were able to connect the different categories to the curriculum they use in the day-to-day practice of the school.

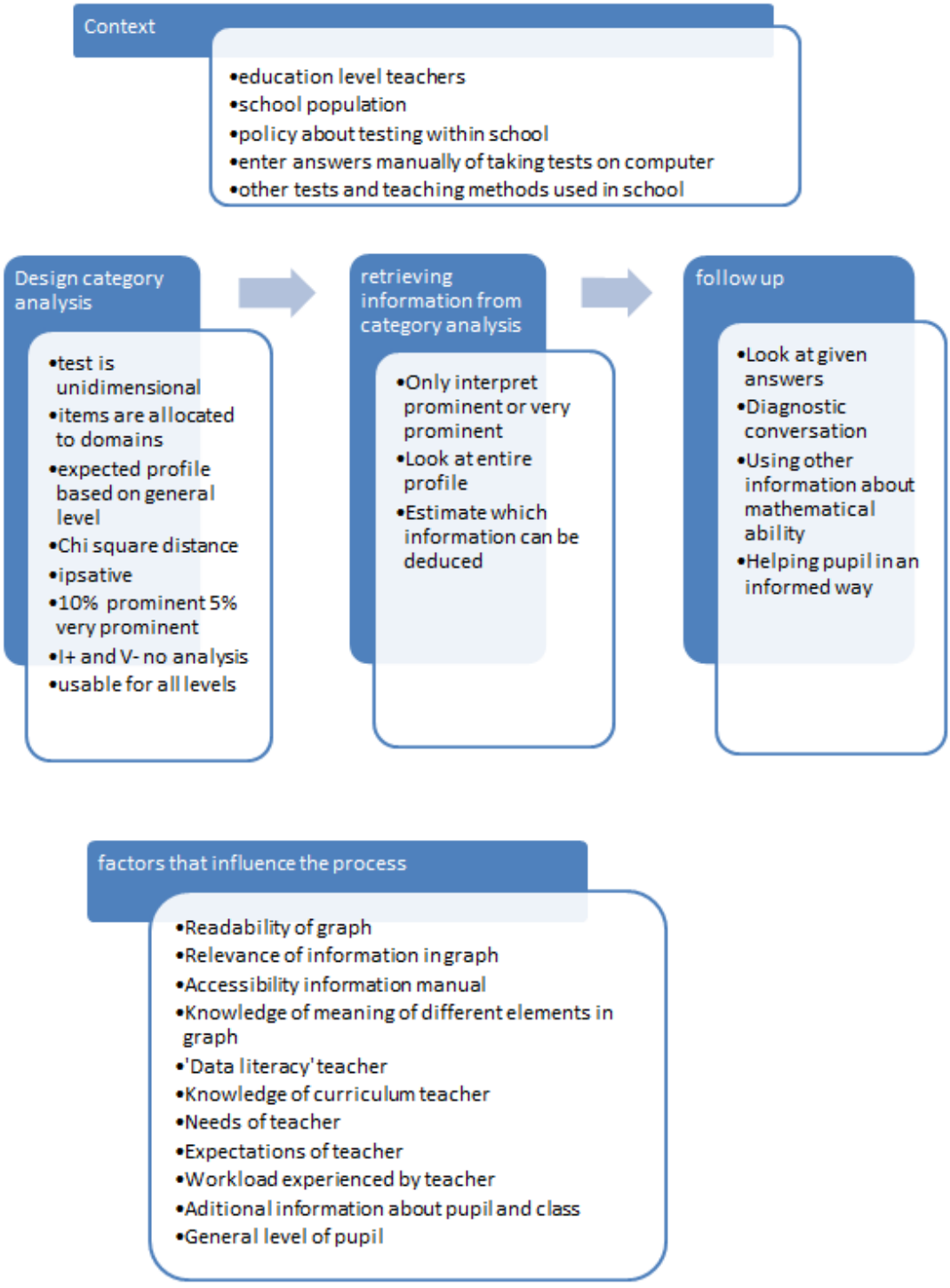


Figure 4. Model for design and use of the CA as deduced from expert interviews

The focus groups showed respondents did not understand all elements of CA and had problems using CA as intended by the Cito experts. Both focus groups interpreted the bars as a distance between the score on the test and an expected score, but the meaning of this expected score was not clear. One focus group reported they saw the baseline of the display, depicting the expected profile, as a ‘norm’: they thought bars below the line indicated insufficient mastery. The second group thought it indicated

the expected growth based on earlier tests. The general concept of CA was not known to the participants of the focus groups. The first focus group mentioned spontaneously they did not read the manual about it, because it was too long. When the researcher explained the meaning of the CA dashboard at the end of the interview, several participants expressed they found it difficult to comprehend.

When interpreting the dashboards the participants in both groups interpreted the bar-graphs whether or not the indication was 'prominent' in an equal way. Percentages of deviation were given a meaning in itself. Participants expressed a concern for certain categories, for example they indicated a score of -8 as a large score which gave reasons for concern, according to them. Both groups mentioned they would like to know the amount of items with a wrong answer because that is understandable information for them, opposed to the percentages presented which were not clear to them.

The participants reported using the findings of CA. In both groups participants indicated the information of CA leads to further investigation to the possible problems of the pupil. However, some participants also indicate the information is used to form groups of children with problems on a certain category, which is not the intended use of Cito because it does not take the general level differences between children into account.

The redesigned dashboards were received with mingled reactions. In general participants liked the clear appearance compared to the original display. When shown displays without numbers, both focus groups indicated they needed numbers to interpret the dashboard. One group did indicate the use of the table, the other indicated they never looked at the table. Both groups had difficulty explaining the meaning of the numbers they indicated using, but expressed they needed them anyway. The use of colour and symbols that indicated 'not prominent', 'prominent', or 'very prominent' stood out, but the use of the colour red provoked strong reactions and was disliked by the participants because they thought it was too negative. In spite of the clear indication of 'not prominent' profiles, the first focus group treated these profiles in the same way as they treated the original dashboards. One participant expressed this in the following way: "Wat mij ook opvalt is dat er staat bij 'alle categorieën', dat het niet opvallend is, terwijl ik dit zie en denk, nou... ik vind het wel opvallend!" [What strikes me is, it says for 'all categories' it is not prominent, but when I see this and think, well… I think this ís prominent!]. The second focus group was also shown a warning text about possible reasons of deviating scores and reacted they could not interpret the alternative dashboards indicated with 'not prominent'.

Both focus groups mentioned they did not like it when the message "Categorieënanalyse niet zinvol vanwege een te hoge/lage score" [category analysis not meaningful due to a score that is too high/low] was shown. When asked, the first focus group stated explicitly, they would like to see it anyway, even though it would not be statistically reliable. The second focus group expressed a wish to be able to see more statistics when they clicked on different elements of the dashboard.

Table 1: results of focusgroups

Aspects of CA	**Views of focus group participants**	
Design CA	***Positive***	***Problem***
Unidimensional design		• One group indicates it is difficult to interpret the categories with different amounts of questions. • No knowledge of item weights.
Items are allocated to domains	• Teachers in both groups give meaning to different domains.	
Expected profile based on general level		• One group interprets level (green line) as a general norm: fail/pass • Concept of comparison to expected profile is not understood (e.g. mastery or growth are perceived).
Chi square distance		• Percentage score deviation is interpreted as a meaningful value. • Different surfaces of bars due to weight is not understood • Wish to see amount of wrong items, which is more meaningful than percentage distance.
Ipsative		• One group thinks all scores below the "norm" is possible.
10% prominent, 5% very prominent		• It is not understood when prominent or very prominent is indicated. • Different contributions of categories to indication is not understood.
I+ and V- no analysis		• Analysis for V- and I+ is missed. One group wants it despite of statistical problems.
Usable for all levels	• Also usable for better achieving pupils.	• One group is surprised to see an indication 'very prominent' for a general level above average.
Retrieving information	***Positive***	***Problem***
Look at entire profile		• Focus on every bar under the line: 'expected profile'. • Concept of expected profile is not understood.
Only interpret 'prominent' or 'very prominent'		• Every profile is interpreted, even though 'not prominent' is spotted.
Estimate which information can be deduced	• Information out of the display is linked to real life situations.	• The second part of the display is not linked to the domain in the first part. • Information in table is difficult to understand and interpret • Scale -25% to 25% is used as a reference.
Follow up	***Positive***	***Problem***
Look at given answers	• Both groups use this.	
Diagnostic conversation	• Mentioned once in one focus group as possible action.	• This is not mentioned often as a follow-up action.
Using other information	• Pupils work in math book. • Earlier Cito test results • 'Cito basisbewerkingen' (one focus group).	
Helping pupil in an informed way	• Participants have the feeling they are better informed to aid the pupil	• In one focusgroup formation of groups with the same problem was mentioned, which does not take the average level into account.

3.3 Survey

3.3.1 Interpretation as intended and knowledge

For the six dashboard interpretation questions and four knowledge questions, a maximum score of 10 points maximum could be gained. The 90% norm-score could be achieved by getting 9 out of ten points. The mean score for all questions was 4.0 points (N=278, SD=2.19) which corresponds to an average percentage score of 40%, which is well below the norm. Only two out of 278 respondents got all ten questions right and one respondent scored nine out of ten, which corresponds with 0.7% of the respondents achieving the norm (N=278). In table 2 and figure 6, all scores per question are displayed and this shows no single question is answered correctly by 90% of the respondents. Overall, interpretation of all dashboards results in a correct percentage of 36.8% (M= 2.21, SD = 1.58).

Especially the questions about interpreting dashboards 1 and 1A (see Appendix C), indicated ‘not prominent’ resulted in low scores. The original display, dashboard 1, was interpreted as intended by 9.0% of the respondents (N = 278, SD = .29) and the alternative display, dashboard 1A, was interpreted as intended by 11.9% of the respondents (N = 278, SD = .32). Answers were indicated as not in accordance with the intentions when respondents indicated they wanted to investigate categories, in spite of the profile being indicated as ‘not prominent’.

Dashboards 2 and 2A (see Appendix C), which showed ‘very prominent’ profiles, were answered according to intention more often, but did not meet the norm. The original dashboard 2 was interpreted as intended by 66% of the respondents (N = 278, SD = .48), the alternative dashboard 2A was answered correct by 51% of the respondents (N = 278, SD = .50). Answers which were indicated as not as intended were mainly caused by indicating the wish to investigate the category ‘meten, tijd en geld’. This domain deviated -7% from the expected score in this example, which is only a small deviation from the expected profile. This category was chosen by respectively 31% and 44% of the respondents for the original and the alternative display.

The third set of dashboards: 3 and 3A (see Appendix C), showed a profile indicated ‘not prominent’ with a sub-category analysis indicated as ‘prominent’. The original dashboard 3, was answered according to intentions by 41% of the respondents (N = 278, SD = .49) and the alternative dashboard 3A was answered according to intentions by 42% of the respondents (N = 278, SD = .50). Questions were rewarded a point for being interpreted as intended, when only bars in the ‘prominent’ sub-profile were selected by the respondents, if the category ‘optellen en aftrekken’ was chosen, this was not indicated as not in accordance with the intended interpretation, because this category had a sub-analysis indicated as ‘prominent’.

The knowledge questions about different aspects of the dashboard had a percentage correct of 44.8% (M=1.79, SD = 1.24). The question about the meaning of the bar graph indicating if the deviation of the profile from the expected profile is ‘(very) prominent’ or ‘not prominent’, was

answered correct by 68% of the respondents, while the other three knowledge questions were answered right between 30% and 43% of the respondents.

Table 2. Mean and percentage correct answers of interpretation and knowledge questions

Question	N	Mean	Percentage correct	SD
score display 1	278	.09	9.0	.287
score display 1A	278	.12	11.9	.324
score display 2	278	.66	65.8	.475
score display 2A	278	.51	51.1	.501
score display 3	278	.41	41.0	.493
score display 3A	278	.42	42.1	.495
knowledge 1	278	.31	30.6	.462
knowledge 2	278	.43	43.2	.496
knowledge 3	278	.38	38.1	.487
knowledge 4	278	.68	67.6	.469
Sum of display scores (max. 6)	278	2.21	36.8	1.58
Sum of knowledge questions (max. 4)	278	1.79	44.8	1.24

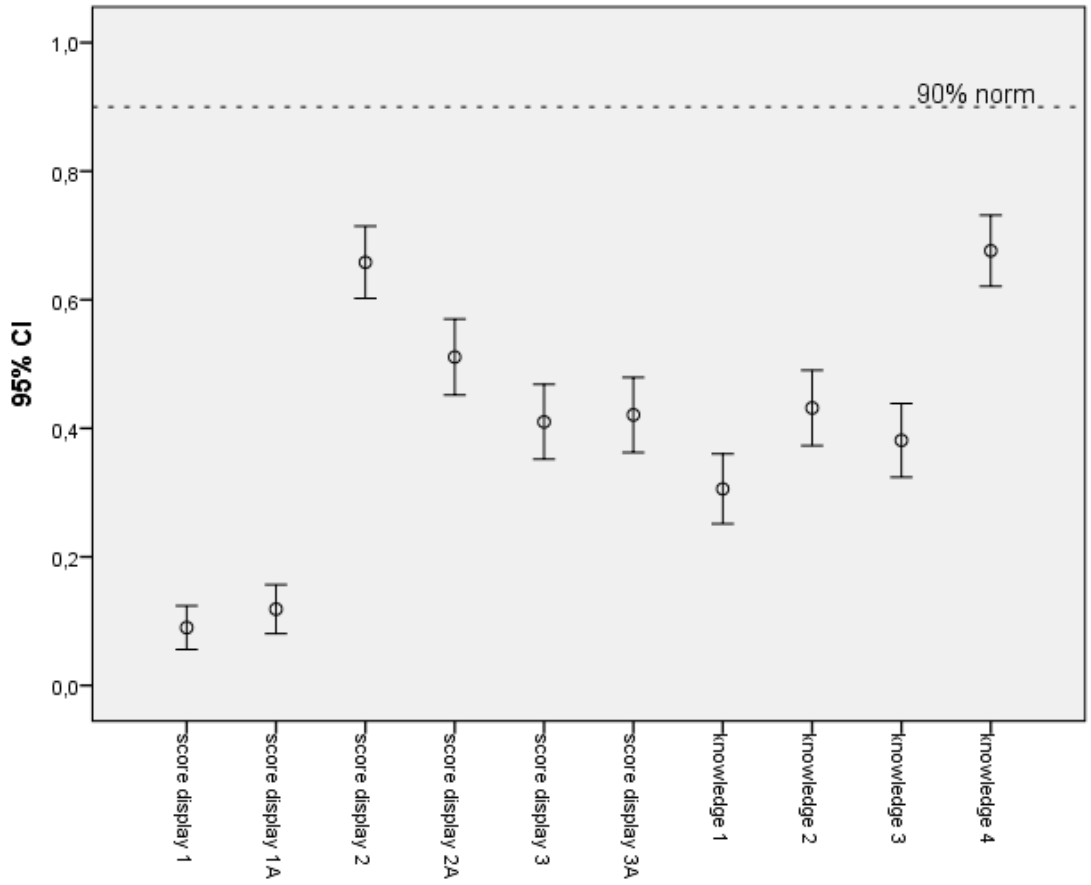


Figure 5. Average proportion correct answers to the display interpretation and knowledge questions

Paired samples T-tests were performed to analyse the differences between the original displays and the alternative displays. The three original displays, dashboards 1, 2 and 3 (see Appendix C) were interpreted as intended by an average of 39% correct (M = 1.16, SD = .52) and the alternative displays, dashboards 1A, 2A and 3A, were interpreted as intended by an average of 35% correct answers (M = 1.05, SD = .53). The score on the original displays is .11 higher than the scores on the alternative scores. This is a significant difference, t (277) = 2.36, p = .019, but it is a small-sized effect, cohen's d = .241.

For the dashboards displaying a profile indicated as 'not prominent', dashboard 1 and 1A (see Appendix C), the percentage correct for the original dashboard (M= .09, SD = .29), was 2.9% lower than the alternative dashboard (M = .12, SD = .32). This difference proved not to be a significant difference, t (277) = -1.80, p = .074.

The original dashboard displaying a 'very prominent' profile, dashboard 2 (see Appendix C), scored 14.7% higher (M = .66, SD = .48) than the alternative display, dashboard 2A (see Appendix C), (M = .51, SD = .50). This is a significant difference, t (276) = 4.86, p < . 001, cohen's d = .30, which is a rather small effect.

The original dashboard with a profile indicated 'not prominent' and a sub-profile indicated as 'prominent', dashboard 3 (see Appendix C), scored 1.1% lower (M = .41, SD = .49) than the alternative display of the same profile, dashboard 3A (see Appendix C), (M = .42, SD = .50), which was not a significant difference, t (277) = -.35, p = .726.

Approximately half of the respondents had been shown a warning message about deviation and chance (n = 125) the other group had not been shown this message (n = 153). To distinguish the difference between the way the displays are interpreted by these groups, an independent t-test was performed. On average the group that saw the message achieved a higher score on the interpretation of the displayed dashboards (M = 2.36, SD = 1.59), than the group that did not see the message (M = 2.08, SD = 1.56). However, the difference is not significant, t (276) = 1.45, p = .149.

An ANOVA was performed to test the effect of function on the total score of the interpretation and knowledge questions. Table 3 shows the scores of different function groups. There was a significant effect of function on the total test scores F(2, 270)= 3.359, p = .036, with a small effect size ω = .13. A Post hoc analysis, using Games Howell, shows there is a significance difference between internal support teachers and managers (average difference = 1.26, p = .023). There are no significant differences between internal support teachers and teachers (average difference = .21, p = .796), or between teachers and managers (average difference = 1.05, p = .106).

Four different groups of experience were distinguished: users who indicated they did not know or ever used CA, users that had used CA incidentally in the past, users who use CA approximately once a year and users who use CA more than once a year. Their scores are displayed in table 4. An ANOVA

was used to distinguish differences between these groups. There was a significant effect of experience on the score, F (3, 277) = 3.095, p = .027, but the effect size is small, ω = .15. Post hoc analysis, using Games Howell, shows there is a significant difference between frequent users and never users (average difference = .826, p = .046).

Table 3. Mean and percentage correct answers by different function groups

User group	n	Mean	Percentage correct	SD
Teacher	59	3.92	39.2	.2.184
Internal support teacher	192	4.13	41.3	2.191
Manager	22	.66	65.8	1.959
Total	273	3.98	39.8	2.191

Table 4. Mean and percentage correct answers by different experience groups for usage CA

User group	n	Mean	Percentage correct	SD
Never used	58	3.53	35.3	1.912
Used in past	44	3.50	35.0	2.085
Use once a year	26	3.85	38.5	2.130
Use frequent	150	4.36	43.6	2.286
Total	278	4.00	43.2	2.192

3.3.2 EFLA

Respondents that had indicated they used CA at least once a year (n= 176) answered questions of the translated EFLA questionnaire, in respect to their general opinion towards CA. The questions of the EFLA questionnaire were scored on a ten point Likert scale, in which ten points is the highest (positive) score. On average, the total score on the EFLA for CA was 6.4. All but one of the scores varied with a mean score between 6.4 and 7.3. Only question four, asking about the ability of the tool for predicting the future learning achievements, was scored with a mean of 5.4. The score on the factor DATA was 6.8, on Awareness & Reflection 6.3 and on Impact 6.2. The scores on each question are displayed in table 5.

Table 5. Mean and standard deviation of the EFLA questionnaire on CA

Items	Mean	SD
Voor de categorieën analyse is het is duidelijk welke gegevens verzameld worden.	6.91	2.20
Voor de categorieën analyse is het duidelijk waarom de gegevens verzameld worden	7.27	2.06
De categorieën analyse maakt me bewust van de huidige leerprestaties van mijn leerling.	7.14	1.95
De categorieën analyse laat me de mogelijke toekomstige leerprestaties van mijn leerlingen voorspellen, op grond van hun (on)veranderde gedrag.	5.35	2.17
De categorieën analyse stimuleert me om te reflecteren op mijn leerkrachtgedrag.	6.96	2.02
De categorieën analyse stimuleert me om mijn leerkrachtgedrag aan te passen als dat nodig is.	7.15	1.95
De categorieën analyse stimuleert me om efficiënter les te geven.	6.43	2.23
De categorieën analyse stimuleert me om effectiever les te geven.	6.64	2.16

4. Discussion and Conclusion

The central question of this thesis was: “Is the ‘Cito category-analysis’ a tool that gives meaningful information to teachers and internal support teachers in Dutch primary education?” Therefore, the intended interpretation of CA, the way users interpret and value CA, and if design enhancements can improve interpretation CA by its users were investigated. This was done by interviewing Cito experts, designing an alternative dashboard, interviewing two focus groups and distributing a survey. In this way qualitative and quantitative data were gathered.

The first research question: “What is the intended interpretation of the ‘Cito category-analysis’ dashboard?”, resulted in a model (figure 5) for the design and use of CA. This model was based on interviews with Cito experts and literature study. The first step, when a user interprets CA, according to the interviewed experts, is to look at the indication ‘(very) prominent’ or ‘not prominent’. When a profile is indicated as ‘not prominent’, users should not interpret the profile any further. When a profile is indicated ‘(very) prominent’, results should be interpreted with care and be further investigated. A diagnostic conversation, or looking at other mathematical work of the pupil is recommended.

The second research question: “Which enhancements or alternatives can be designed, based on expert opinion, users’ views and literature, to help users interpret the ‘Cito category-analysis’ dashboard in an appropriate way?”, led to an alternative dashboard. This alternative dashboard (figure 4, right) was created, based on the aims of use of CA expressed by the Cito experts and the design principles of Kosslyn (2006). The general aim of the redesign was focused on a more clear distinction between ‘not prominent’, ‘prominent’ and ‘very prominent’ profiles and to show only the information that gives meaningful information to the reader. The alternative dashboards were appreciated by the focus groups for their clear appearance. Focus group participants had strong feelings about the use of the colour red, because of the negative association, but the indication ‘not prominent’, ‘prominent’ or ‘very prominent’ caught their eye immediately. Focus group participants generally expressed they wanted to see numbers, although they could not explain the meaning of different numbers when asked.

The third research question was: “How is the ‘Cito category-analysis’ evaluated by experienced and intermediate users?” The evaluation of CA by users in two focus groups was moderately positive. Participants reported they valued CA and used it frequently, in spite of manually entering data into the Cito LOVS computer program, which they found time consuming. This was supported by the EFLA questionnaire. The scores on different factors were moderately positive and indicate respondents know to a certain extent what data are used and why (score 6.8), are being made more aware of the achievements of their pupils and their own teaching behaviour (score 6.3) and find CA influences their future teaching moderately (score 6.2). The question: CA makes me forecast my students possible future learning situation given their (un)changed behaviour, was rated lower and nearly neutral (score 5.4). This is understandable since this is not the intention of Cito LOVS tests nor of CA. Although the

participants of the focus groups reported they used CA frequently, they also indicated they found it hard to understand the real meaning of the analysis.

The fourth research question was: “Is the actual interpretation of the ‘Cito category-analysis’ dashboard by teachers and internal support teachers in accordance with the intended interpretation?” The actual interpretation of the CA dashboard by users seemed to differ dramatically from the intentions. During the focus group interviews it became clear that users had difficulties interpreting the CA dashboard. Participants of the focus groups noticed if profiles were indicated as ‘(very) prominent’ or ‘not prominent’ profiles, but interpreted all profiles anyway. The problems, as seen in the focus groups, matched with the results of the survey data. Only 0.7% of the respondents managed to score the questions to match the 90% score-norm, set by a Cito expert. The mean score of 37% interpretation and knowledge questions answered correctly according to intentions of Cito experts, raises severe concerns about the way professionals using Cito LVS tests are able to interpret the CA dashboard.

The question about the original dashboard displaying a profile which is indicated ‘not prominent’ was interpreted in accordance to the intentions by only 9% of the respondents. These profiles make up approximately 90% of all CA dashboards. Users seemed to trust their own judgement, based on how they interpret the bar graph, better than the statistical interpretation Cito gives. Which was illustrated by a participant: “Wat mij ook opvalt is dat er staat bij alle categorieën dat het niet opvallend is, terwijl ik dit zie en denk, nou... ik vind het wel opvallend!” [What strikes me is, it says for all categories it is not prominent, but when I see this and think, well… I think this is prominent!] The fact that these profiles were interpreted despite of the indication ‘not prominent’, suggests the CA dashboards are overused by the vast majority of the users. Although graphics help users of dashboards to understand the deeper structure of complex data (Hattie & Brown, 2007), they give highly salient and persuasive information and may lead to overuse of numbers (Zapata-Rivera & Zwick, 2011), which seems to be the case here.

Users of CA dashboards seemed to have limited and sometimes incorrect knowledge about several vital aspects of the dashboard. In the survey only 45% of the knowledge questions was answered correctly. In the focus groups misconceptions about CA elements also emerged. One focus group interpreted the expected profile, based on the total test score of the pupil, as a norm which indicates if the pupil masters a category. On top of this, reported numbers seemed to contribute to misconceptions: percentage deviations from the expected profiles were interpreted as values in itself, even though this percentage has no particular meaning and respondents were not able to explain what the meaning of the numbers was.

The possible follow up actions, as mentioned in the focus groups, were more in accordance with the intentions of Cito experts, but not entirely. The mathematical domains, that are distinguished in CA, are meaningful for the educators in the focus groups. Investigating given answers of prominent

profiles was mentioned, as well as the use of other work of pupils on math. The use of diagnostic conversations was only mentioned once and seemed not to be used often, although it is recommended by the Cito experts as an important tool. In one focus group it was mentioned that pupils were assigned into extra instruction groups on domains with negative scores in CA. This is not the intended use of CA, because you have to take the general level of the pupil into account, which was not mentioned in this case.

The fifth research question: "Is the interpretation of the alternative dashboard by teachers and internal support teachers more in accordance with the intended interpretation?", was investigated by focus groups and a questionnaire. The redesigned dashboard did not lead to better results. One focus group decided they could not interpret the profiles indicated 'not prominent' when they were shown the alternative dashboard after they had read a message about chance. This effect was not seen in the survey where there was a significant lower score on the interpretation of the alternative dashboard and the warning message on chance caused no significant difference in interpretation. The lower score for the redesigned dashboard was caused by respondents indicating a wish to investigate the cause of a small deviating category in the 'very prominent' profile. Possibly, this effect was caused by the fact that no changes to the surface were made in the alternative display and the bar looked bigger as a consequence

The sixth research question was: "Is there a difference in interpretation of the 'Cito Category Analysis' dashboard and the alternative dashboard between experienced users, intermediate users and new users, or between teachers, internal support teachers and school leaders?" There was a significant positive effect of experience on interpreting CA, but the effect size was very small and only between the group of no users and experienced users a significant effect could be distinguished. No significant difference was discovered between different user groups interpreting CA. This result is not consistent with the findings of Van der Kleij & Eggen (2013), who found internal support teachers seemed to be better able to interpret Cito score reports than teachers and managers. This different outcome may be caused by the groups of respondents that were invited: Cito portal administrators and a group of active users of Cito products. Teachers in this group might be more experienced with Cito test reports than average.

The seventh and final research question was: "What guidelines can be deduced for a future redesign of the "Cito Category Analysis" in particular and graphics in pupil monitoring systems for educators in general?" This study showed users of CA had great difficulties understanding the general concept of CA and the way this was shown in a dashboard. The use of a statistical concept such as significance, even if it is translated in easier terms as 'prominent' or 'not prominent', proved to be difficult for users when combined with graphics that seem to give highly salient information. It is questionable if data should be presented at all in case of 'not prominent' profiles. Even in case of presenting a profile that is indicated as 'very prominent', the interpretation by users seems to lead to

over-signalling of suspected problems, when presented graphically. Alternatively, tables or text can be used instead of graphics, because they are relatively 'neutral' to interpretation (Zapata-Rivera & Zwick, 2011). Further investigation and field testing of possible designs for CA without using graphics is recommended.

The way respondents were approached gives reason to believe they do not represent all teachers and internal support teachers in The Netherlands. There is reason to believe the participating respondents, especially the participating teachers, may have more interest and experience in using Cito tests than the average group. Also, the total group contains a majority of internal support teachers and the average age of the respondents is higher than the average age of the educational professionals working in primary education. Consequently, a higher experience in education in general and interpreting test reports specifically, can be presumed. This might even emphasise the concerning results on knowledge and interpretation of CA dashboards, as shown in this study.

This thesis inquires respondents about interpretation of CA dashboards and interprets the answers as correct or incorrect. As interpretation is not absolute and open to discussion, this way of treating the answers may also be discussed. Especially the criterion 'not prominent' that leads, in this thesis, only to the conclusion no inferences should be made from the CA dashboard, can be debated. As stated before, the user manual of Cito LOVS (Cito, 2018) encourages the user also to look at 'nearly prominent' dashboards, because this might give valuable information that might otherwise be missed. Looking at nearly 'prominent' profiles has the advantage that the user will not miss possibly important information because of dogmatic 'all-or-nothing thinking', which distinguishes nearly identical deviations possibly causing type 2 error (Field, 2013). On the other hand, only looking at '(very) prominent' profiles reduces the chance of type 1 error, in which the hypothesis: the ability for maths indicates a unidimensional continuum, is rejected on false grounds (Field, 2013). In this study the expert opinion was taken as a guideline for the intended use of CA, because it is the most clear guideline to investigate intended use and it protects the user for overuse of data on false grounds.

When score reports and educational dashboards are used in education, the correct interpretation of the presented results by the readers is of the utmost importance and effects the validity of usage of the test report (Gotch & Roduta Roberts, 2018; J. Hattie, 2009; O'Leary et al., 2017). The way users interpret the CA dashboard raises serious concerns about this validity of usage. This investigation suggests the use of CA leads to overuse of numbers and might even lead to give wrong or unnecessary remediation to the test taker (Hambleton & Zenisky, 2013; Monaghan, 2006).

In this investigation respondents gave rather positive reactions to CA in the EFLA questionnaire and in the focus groups. This is striking, since their conception of the CA dashboard is not in compliance with the intentions of the Cito experts. This is in line with reported low data literacy of teaching professionals (e.g. Mandinach, 2012; Mandinach & Gummer, 2013; Schildkamp, Karbautzki, & Vanhoof, 2014; Vanhoof, Verhaeghe, Verhaeghe, Valcke, & Van Petegem, 2011). Overuse of

presented data and misconceptions seem not to be sensed by teaching staff. And although the current state of low data literacy due to insufficient education (Ledoux et al., 2009) is recognised by the author of this thesis, this gives developers a big responsibility in fitting test reports in relation to the intended users and their data literacy.

This investigation supports Ryan (2006) who states test reports should be field tested, comparable to the testing of test/quiz/survey items, when designed and used in the initial phase. Findings of this study also indicate further research to make data more accessible to teaching staff in general and teachers in primary education in particular, is needed.

5. References

Beiro, L. F., & Ramaekers, M. (2016). Steeds meer vrouwen voor de klas in het basisonderwijs. *CBS Sociaaleconomische trends, 08.* Retrieved from: https://www.cbs.nl/-/media/_pdf/2016/40/steeds-meer-vrouwen-in-het-onderwijs.pdf

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5-25.

BTC Media Test BV. (2018). *Op zoek naar de wow-factor; Onderzoek naar de klantreis en conceptvalidatie LVS 4.0 PO* (report number: R-4153, D1). Amsterdam.

Cavalluzzo, L., Geraghty, T. M., Steele, J. L., Holian, L., Jenkins, F., Alexander, J. M., & Yamasaki, K. Y. (2014). " Using Data" to Inform Decisions: How Teachers Use Data to Inform Practice and Improve Student Performance in Mathematics. Results from a Randomized Experiment of Program Efficacy. *CNA Corporation*.

Centraal Bureau voor de Statistiek. (2019a, May). Werkzame beroepsbevolking; beroep [Data file]. Retrieved from https://opendata.cbs.nl/statline/#/CBS/nl/dataset/82808NED/table?dl=AF86

Centraal Bureau voor de Statistiek. (2019b, June). *Regionale kerncijfers Nederland* [Data file]. Retrieved from https://opendata.cbs.nl/statline/#/CBS/nl/dataset/70072ned/table?ts=1562413050676

Cito. (2018). *Computerprogramma LOVS gebruikershandleiding* (4.11b ed.). Arnhem: Author.

Creswell, J. W. (2014). *Educational Research: Planning, Conducting and Evaluating Quantitative and Qualitative Research* (Fourth ed.). Essex: Pearson.

Davenport, T. H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Boston: Harvard Business Press.

Faria, A.-M., Heppen, J., Li, Y., Stachel, S., Jones, W., Sawyer, K., . . . Lewis, S. (2012). *Charting Success: Data Use and Student Achievement in Urban Schools.* Washington: Council of the Great City Schools. Retrieved from https://files.eric.ed.gov/fulltext/ED536748.pdf

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*(4th edition). London: Sage.

Goodman, D. P., & Hambleton, R. K. (2004). Student Test Score Reports and Interpretive Guides: Review of Current Practices and Suggestions for Future Research. *Applied Measurement in Education, 17*(2), 145-220.

Gotch, C. M., & Roduta Roberts, M. (2018). A Review of Recent Research on Individual-Level Score Reports. *Educational Measurement: Issues and Practice, 37*(3), 46-54.

Hambleton, R., & Zenisky, A. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. *APA handbook of testing and assessment in psychology, 3*, 479-494.

Hattie, J. A. (2009). Visibly learning from reports: The validity of score reports. *Online Educational Research Journal*.

Hattie, J. A., & Brown, G. T. (2007). Technology for school-based assessment and assessment for learning: Development principles from New Zealand. *Journal of Educational Technology Systems, 36*(2), 189-201.

Hattie, J. A., Brown, G. T., & Keegan, P. J. (2003). A National Teacher-Managed, Curriculum-Based Assessment System. *International Journal of Learning, 10*.

Hop, M., & Engelen, R. (2017). Wetenschappelijke verantwoording Rekenen-Wiskunde 3.0 voor groep 6. In Cito (Ed.). Arnhem.

Hopster-den Otter, D., Wools, S., Eggen, T. J. H. M., & Veldkamp, B. P. (2017). Formative use of test results: A user's perspective. *Studies in Educational Evaluation, 52*, 12-23. doi:10.1016/j.stueduc.2016.11.002

Inspectie van het Onderwijs. (2010). *Opbrengstgericht werken in het basisonderwijs.* Utrecht: Author.

Inspectie van het Onderwijs. (2018). *Rapport-De staat van het onderwijs-onderwijsverslag over 2016-2017*. Utrecht: Author.

Janssen, J. & Hickendorff, M. (2008). Categorieënanalyse bij de LOVS-toetsen rekenen-wiskunde. Arnhem: Cito/Leiden: Universiteit Leiden. Retrieved from: https://anzdoc.com/categorienanalyse-bij-de-lovstoetsen.html

Janssen, J., Verhelst, N., Engelen, R. & Scheltens, F. (2010). *Wetenschappelijke verantwoording van de toetsen LOVS Rekenen-Wiskunde voor groep 3 tot en met 8.* Cito.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educational Researcher, 33*(7), 14-26. doi:10.3102/0013189X033007014

Kosslyn, S. M. (2006). *Graph design for the eye and mind*: New York, USA: Oxford University Press.

Ledoux, G., Blok, H., Boogaard, M., & Krüger, M. (2009). *Opbrengstgericht werken: over de waarde van meetgestuurd onderwijs*: Amsterdam SCO-Kohnstamm Instituut 9789068138702.

Mandinach, E. B. (2012). A Perfect Time for Data Use: Using Data-Driven Decision Making to Inform Practice. *Educational Psychologist, 47*(2), 71-85.

Mandinach, E. B., & Gummer, E. S. (2013). A Systemic View of Implementing Data Literacy in Educator Preparation. *Educational Researcher, 42*(1), 30-37. doi:10.3102/0013189X12459803

Meijer, J., Ledoux, G., & Elshof, D. (2011). *Gebruikersvriendelijke leerlingvolgsystemen in het primair onderwijs*: Amsterdam SCO-Kohnstamm Instituut.

Meijer, P. C., Verloop, N., & Beijaard, D. (2001). Similarities and differences in teachers' practical knowledge about teaching reading comprehension. *The Journal of Educational Research, 94*(3), 171-184.

Monaghan, W. (2006). The Facts About Subscores. *R&D Connections*. Princeton, New Jersey: Educational Testing Service. Retrieved from: https://www.ets.org/Media/Research/pdf/RD_Connections4.pdf

O'Leary, T. M., Hattie, J. A., & Griffin, P. (2017). Actual interpretations and use of scores as aspects of validity. *Educational Measurement: Issues and Practice, 36*(2), 16-23.

Picciano, A. G. (2012). The evolution of big data and learning analytics in American higher education. *Journal of Asynchronous Learning Networks, 16*(3), 9-20.

Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. *Handbook of test development*, 677-710.

Scheffel, M., Drachsler, H., Toisoul, C., Ternier, S., & Specht, M. (2017). *The proof of the pudding: examining validity and reliability of the evaluation framework for learning analytics.* Paper presented at the European Conference on Technology Enhanced Learning. Retrieved from: https://link.springer.com/chapter/10.1007/978-3-319-66610-5_15

Schildkamp, K., Karbautzki, L., & Vanhoof, J. (2014). Exploring data use practices around Europe: Identifying enablers and barriers. *Studies in Educational Evaluation, 42*, 15-24.

Schildkamp, K., & Kuiper, W. (2010a). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education, 26*(3), 482-496. doi:10.1016/j.tate.2009.06.007

Schildkamp, K., & Kuiper, W. (2010b). Data-Informed Curriculum Reform: Which Data, What Purposes, and Promoting and Hindering Factors. *Teaching and Teacher Education: An International Journal of Research and Studies, 26*(3), 482-496.

Schwendimann, B. A., Rodriguez-Triana, M. J., Vozniuk, A., Prieto, L. P., Boroujeni, M. S., Holzer, A., . . . Dillenbourg, P. (2017). Perceiving learning at a glance: A systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies, 10*(1), 30-41.

Shah, P., Freedman, E. G., & Vekiri, I. (2005). The Comprehension of Quantitative Information in Graphical Displays. In A. Miyake & P. Shah (Eds.), *The Cambridge Handbook of Visuospatial Thinking* (pp. 426-476). Cambridge: Cambridge University Press.

Staman, L., Visscher, A. J., & Luyten, H. (2014). The effects of professional development on the attitudes, knowledge and skills for data-driven decision making. Studies in Educational Evaluation, 42, 79-90.

Tempelaar, D. T., Heck, A., Cuypers, H., van der Kooij, H., & van de Vrie, E. (2013). *Formative assessment and learning analytics.* Paper presented at the Proceedings of the third international conference on learning analytics and knowledge.

Twing, J. S. (2008). *Score reporting, off-the-shelf assessments and NCLB: Truly an unholy trinity.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Van der Kleij, F. M. (2013). *Computer-based feedback in formative assessment.* (PhD), University of Twente, Enschede.

Van der Kleij, F. M., & Eggen, T. (2013). Interpretation of the score reports from the Computer Program LOVS by teachers, internal support teachers and principals. *Studies in Educational Evaluation, 39*(3), 144-152.

Van der Kleij, F. M., Eggen, T. J., & Engelen, R. J. (2014). Towards valid score reports in the Computer Program LOVS: A redesign study. *Studies in Educational Evaluation, 43,* 24-39.

Vanhoof, J., Verhaeghe, G., Verhaeghe, J. P., Valcke, M., & Van Petegem, P. (2011). The influence of competences and support on school performance feedback use. *Educational studies, 37*(2), 141-154.

Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. L. (2013). Learning Analytics Dashboard Applications. *American Behavioral Scientist, 57*(10), 1500-1509. doi:10.1177/0002764213479363

Verhelst, N. (2007). *Profielanalyse met Item Respons Theorie*. Arnhem: Cito.

Zapata-Rivera, D., & Zwick, R. (2011). Improving Test Score Reporting: Perspectives from the ETS Score Reporting Conference. Research Report. ETS RR-11-45. *ETS Research Report Series*.

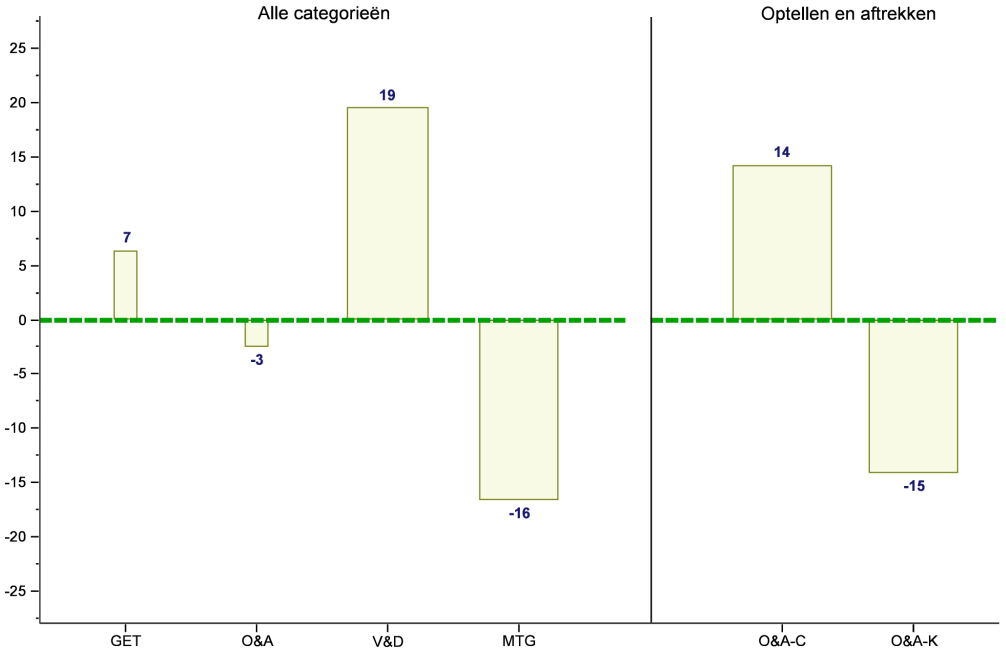
Zapata-Rivera, D., Zwick, R., & Vezzu, M. (2016). Exploring the Effectiveness of a Measurement Error Tutorial in Helping Teachers Understand Score Report Results. *Educational Assessment, 21*(3), 215-229. doi:10.1080/10627197.2016.1202110

Appendix A – Dashboards shown to focus groups

Citogroep - DEMO versie Arnhem

Categorieënanalyse leerling

Leerling: **Safina Krausemann** (5 - 5A)
Toets - taak: **Rekenen-Wiskunde 3.0 - M5**



Afnamedatum: 25-01-2018
Score / Vaard. Niv.: 37 / 194 IV
Alle categorieën: 0 Z Niet opvallend
Optellen en aftrekken: 0 Z Niet opvallend

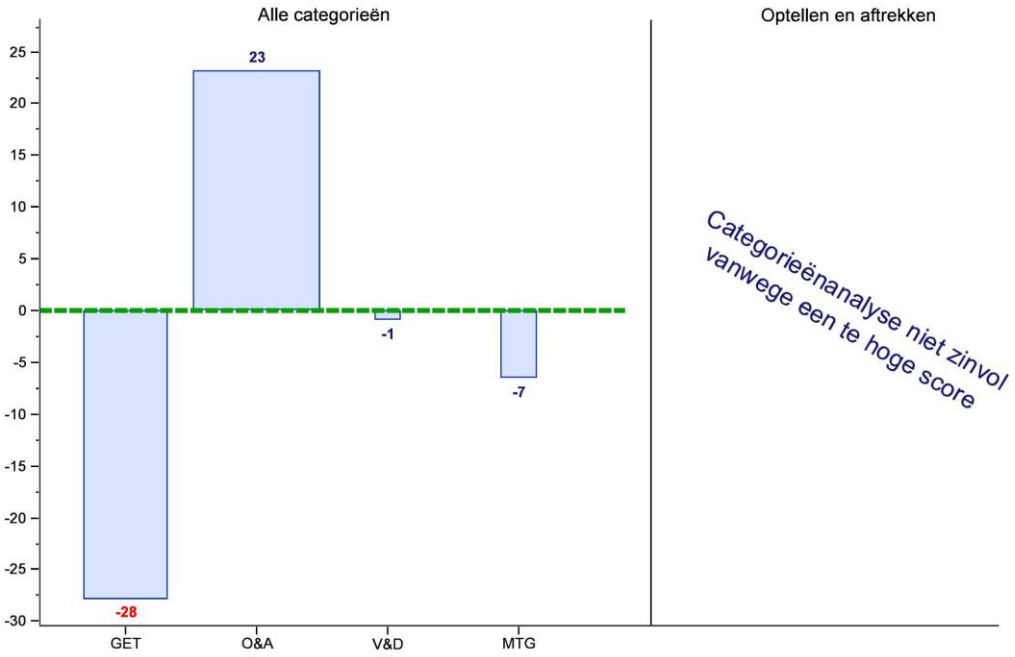
Categorie label	Categorie	Aantal opgaven	% Score geobserveerd	% Score verwacht	% Score afwijking
GET	Getallen en getalrelaties	10	67	60	+7
O&A	Optellen en aftrekken	24	63	66	-3
V&D	Vermenigvuldigen en delen	12	83	64	+19
MTG	Meten, tijd en geld	14	42	58	-16
O&A-C	Optellen en aftrekken - Context	12	81	67	+14
O&A-K	Optellen en aftrekken - Kaal	12	44	59	-15

Appendix A –display pupil 1

Citogroep - DEMO versie Arnhem

Categorieënanalyse leerling

Leerling: **Denise Raven** (5 - 5A)
Toets - taak: **Rekenen-Wiskunde 3.0 - E5-digi**



Afnamedatum: 13-06-2018
Score / Vaard. Niv.: 63 / 222 II
Alle categorieën: O Z Zeer opvallend
Optellen en aftrekken: -

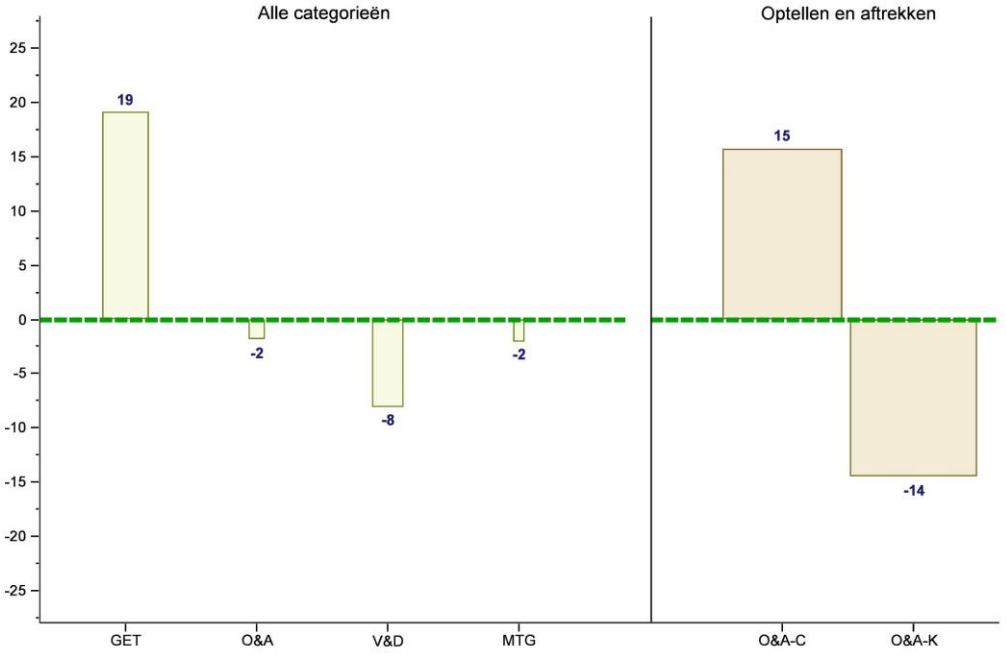
Categorie label	Categorie	Aantal opgaven	% Score geobserveerd	% Score verwacht	% Score afwijking
GET	Getallen en getalrelaties	12	48	76	-28
O&A	Optellen en aftrekken	24	100	77	+23
V&D	Vermenigvuldigen en delen	24	76	77	-1
MTG	Meten, tijd en geld	24	69	76	-7

Appendix A –display pupil 2

Citogroep - DEMO versie Arnhem

Categorieënanalyse leerling

Leerling: **Bruno Haynes** (5 - 5B)
Toets - taak: **Rekenen-Wiskunde 3.0 - M5-digi**



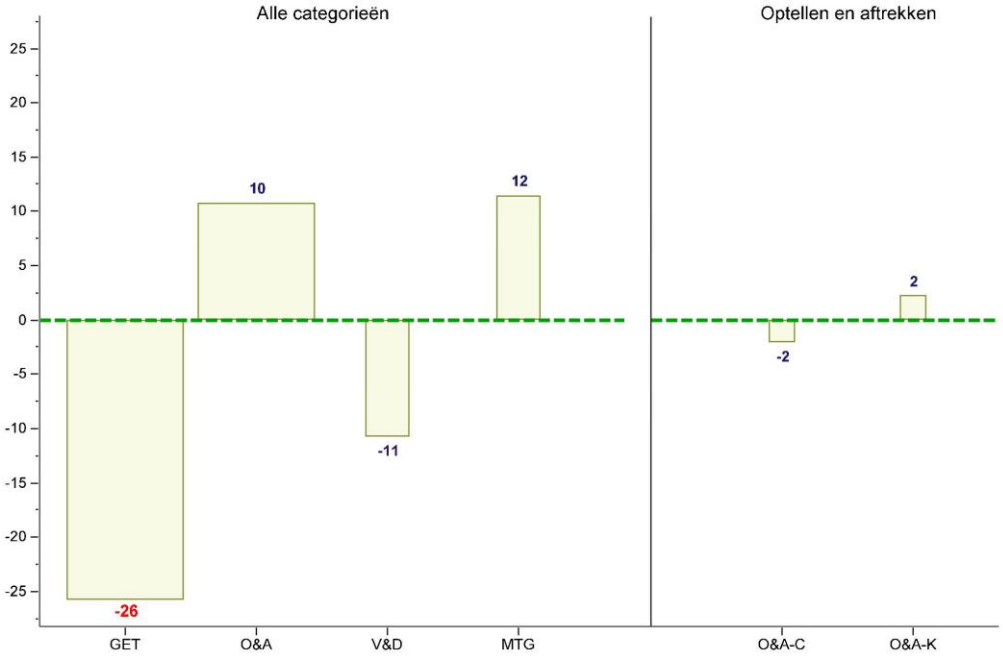
Afnamedatum: 23-01-2018
Score / Vaard. Niv.: 45 / 210 II
Alle categorieën: O Z Niet opvallend
Optellen en aftrekken: O Z Opvallend

Categorie label	Categorie	Aantal opgaven	% Score geobserveerd	% Score verwacht	% Score afwijking
GET	Getallen en getalrelaties	10	88	69	+19
O&A	Optellen en aftrekken	24	76	78	-2
V&D	Vermenigvuldigen en delen	12	67	75	-8
MTG	Meten, tijd en geld	14	75	77	-2
O&A-C	Optellen en aftrekken - Context	12	90	75	+15
O&A-K	Optellen en aftrekken - Kaal	12	63	77	-14

Appendix A – display pupil 3

Categorieënanalyse leerling

Leerling: **Halil Croes** (4 - 4A)
Toets - taak: **Rekenen-Wiskunde 3.0 - M4-digi**



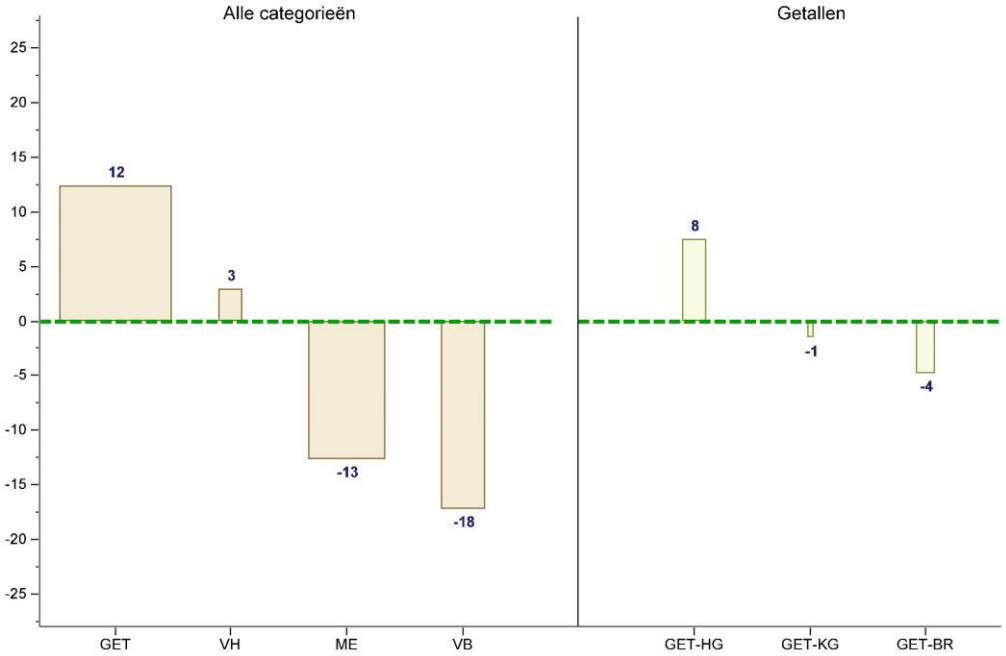
Afnamedatum: 25-01-2018
Score / Vaard. Niv.: 28 / 138 V
Alle categorieën: 0 Z Niet opvallend
Optellen en aftrekken: 0 Z Niet opvallend

Categorie label	Categorie	Aantal opgaven	% Score geobserveerd	% Score verwacht	% Score afwijking
GET	Getallen en getalrelaties	12	34	60	-26
O&A	Optellen en aftrekken	24	57	47	+10
V&D	Vermenigvuldigen en delen	10	35	46	-11
MTG	Meten, tijd en geld	10	53	41	+12
O&A-C	Optellen en aftrekken - Context	12	58	60	-2
O&A-K	Optellen en aftrekken - Kaal	12	57	55	+2

Appendix A –display pupil 4

Categorieënanalyse leerling

Leerling: **Rosario Bemelmans** (6 - 6A)
Toets - taak: **Rekenen-Wiskunde 3.0 - E7-digi**



Afnamedatum: 12-06-2019
Score / Vaard. Niv.: 68 / 271 II
Alle categorieën: 0 Z Opvallend
Getallen: 0 Z Niet opvallend

Categorie label	Categorie	Aantal opgaven	% Score geobserveerd	% Score verwacht	% Score afwijking
GET	Getallen	40	85	73	+12
VH	Verhoudingen	16	79	76	+3
ME	Meten	28	55	68	-13
VB	Verbanden	12	55	73	-18
GET-HG	Getallen - Hele getallen	13	93	85	+8
GET-KG	Getallen - Kommagetallen	14	84	85	-1
GET-BR	Getallen - Breuken	13	80	84	-4

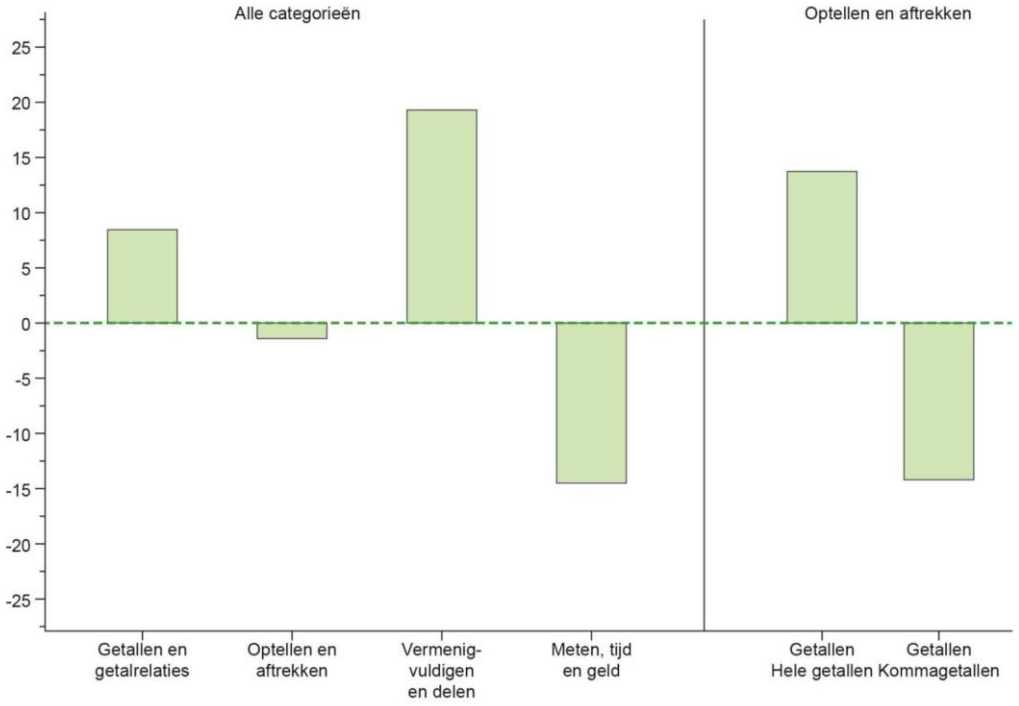
Appendix A – display pupil 5

Citogroep - DEMO versie Arnhem

Categorieënanalyse leerling

Leerling: **Safina Krausemann** (5 - 5A)
Toets - taak **Rekenen - Wiskunde 3.0 - M5**
Afnamedatum: 12-01-2019
Score / Vaard. Niv.: 37 / 194 **IV**

Alle categorieën:	niet opvallend
Optellen en aftrekken:	niet opvallend



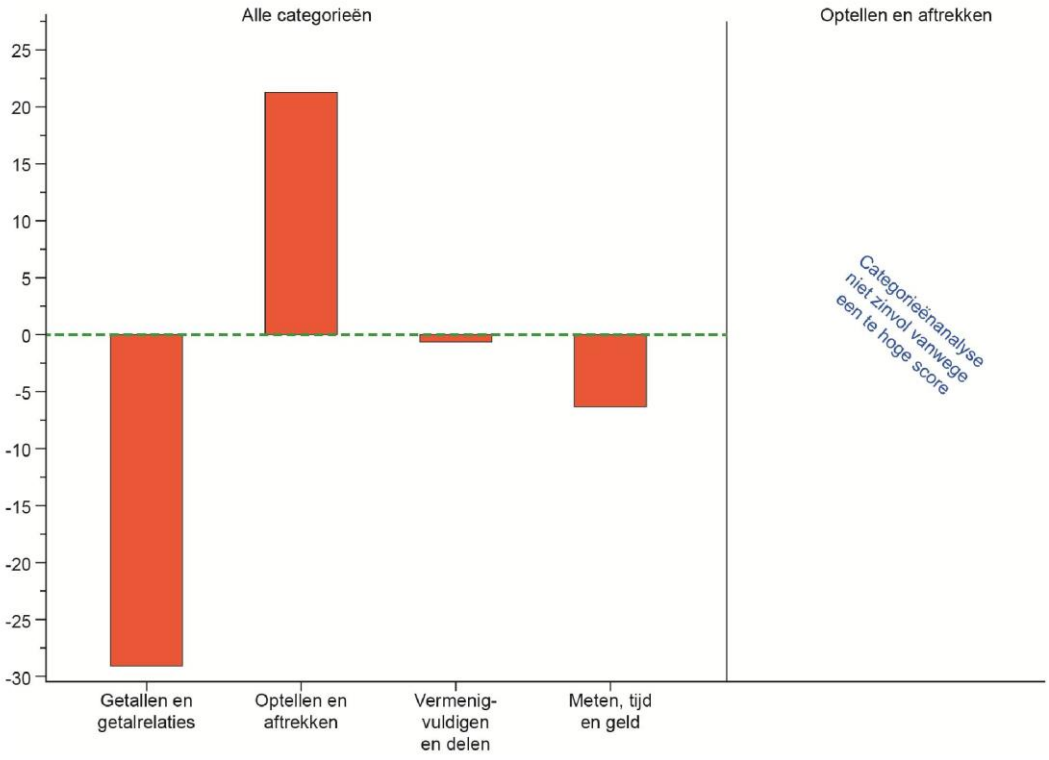
Appendix A - alternative display pupil 1

Citogroep - DEMO versie Arnhem

Categorieënanalyse leerling

Leerling: **Denise Raven** (5 - 5A)
Toets - taak **Rekenen - Wiskunde 3.0 - E5-digi**
Afnamedatum: 12-06-2019
Score / Vaard. Niv.: 63 / 222 II

Alle categorieën: zeer opvallend
Optellen en aftrekken: -



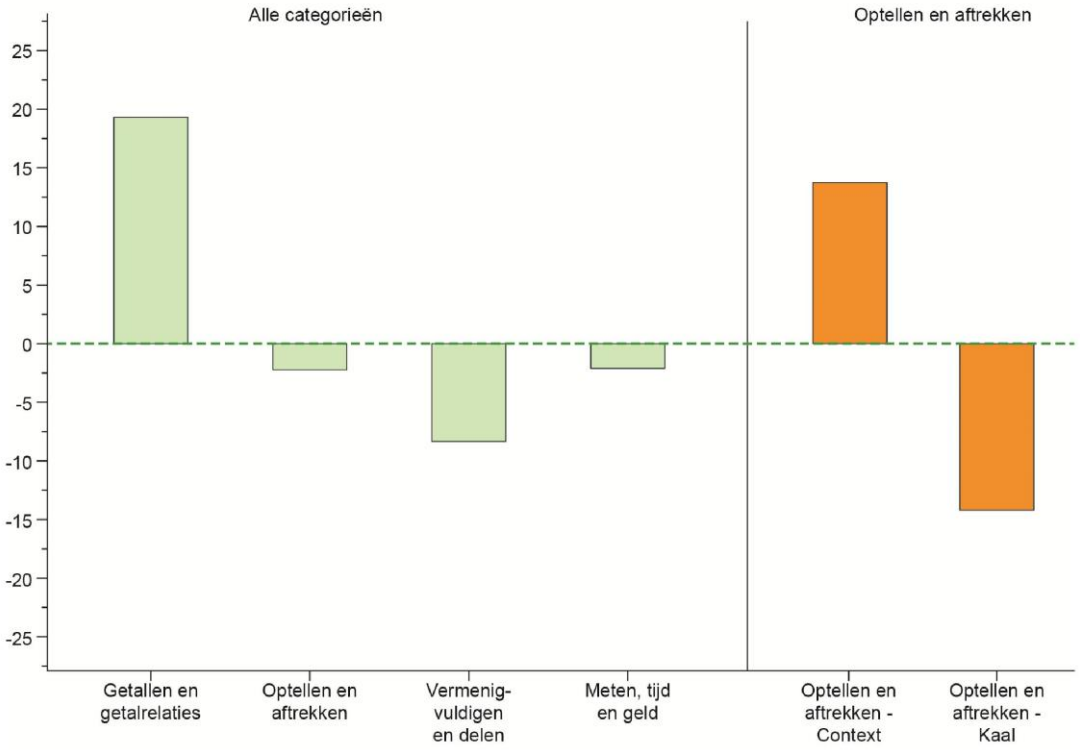
Appendix A - alternative display pupil 2

Citogroep - DEMO versie Arnhem

Categorieënanalyse leerling

Leerling: **Bruno Haynes** (5 - 5B)
Toets - taak **Rekenen - Wiskunde 3.0 - M5-digi**
Afnamedatum: 12-01-2019
Score / Vaard. Niv.: 45 / 210 II

Alle categorieën:	niet opvallend
Optellen en aftrekken:	opvallend



Appendix A - alternative display pupil 3

Citogroep - DEMO versie Arnhem

Categorieënanalyse leerling

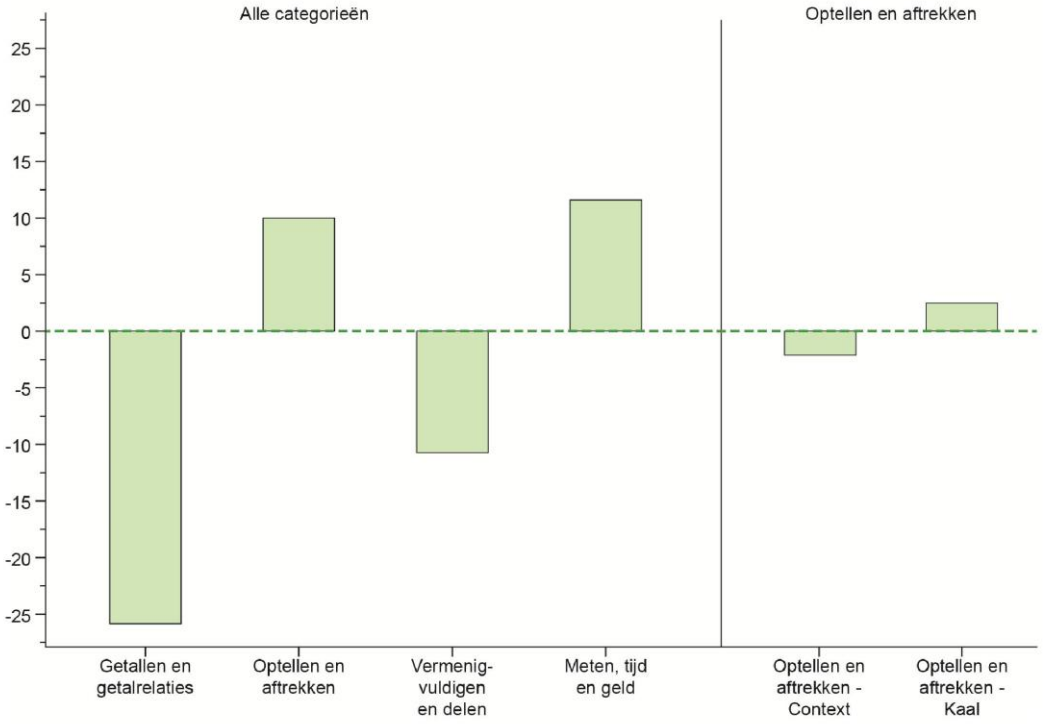
Leerling: **Halil Croes (4 - 4A)**

Toets - taak **Rekenen - Wiskunde 3.0 - M4-digi**

Afnamedatum: 12-01-2019

Score / Vaard. Niv.: 28 / 138 V

Alle categorieën:	niet opvallend
Optellen en aftrekken:	niet opvallend



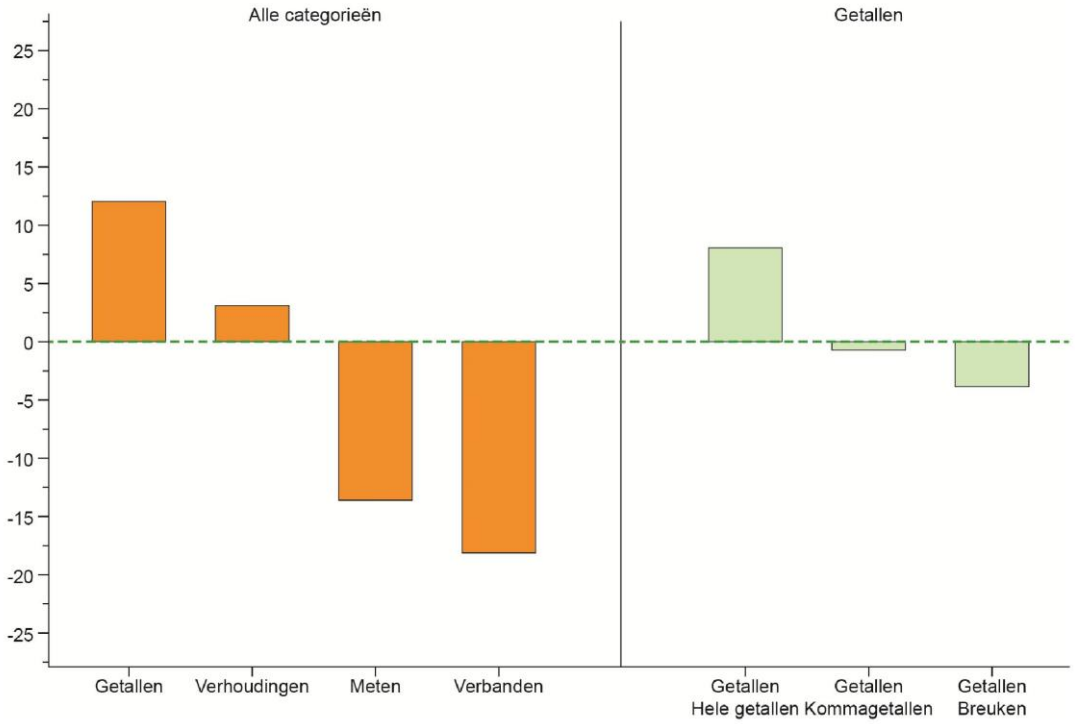
Appendix A - alternative display pupil 4

Citogroep - DEMO versie Arnhem

Categorieënanalyse leerling

Leerling: **Rosario Bemelmans** (6 - 6A)
Toets - taak **Rekenen - Wiskunde 3.0 - E7-digi**
Afnamedatum: 12-06-2019
Score / Vaard. Niv.: 68 / 271 II

Alle categorieën: opvallend
Getallen: niet opvallend



Appendix A - alternative display pupil 5

Appendix B – Protocol focus groups

Willen jullie aangeven wat je functie is, hoe lang je in het onderwijs werkt en wat je ervaring met de categorieën analyse is?

Ik ga zo verschillende weergaven van de categorieën analyse laten zien. Ik wil jullie vragen te vertellen waar je naar kijkt en wat dit voor je betekent. Er is geen goed of fout antwoord en je kunt de grafieken verschillende interpreteren. Als je op elkaar wil reageren, dan is dat prima!

Wat zou je doen als dit één van jouw leerlingen was? Wat zouden je vervolgacties zijn? (zelfde vraag voor elke van de 10 dashboards)

Dit waren de weergaven zoals ze nu worden gegeven door Cito-LOVS. Ik wil jullie nu vragen om te kijken naar de volgende alternatieve weergaven van de categorieën analyse en weer te vertellen waar je naar kijkt en wat dit voor je betekent. Ook hier zijn verschillende interpretaties van de grafiek mogelijk.

Wat zou je doen als dit één van jouw leerlingen was? Wat zouden je vervolgacties zijn?

Wat valt je op aan de originele en alternatieve afbeeldingen? Wat is duidelijk en wat is minder duidelijk/verwarrend.

Sommige grafieken zijn aangegeven als “niet opvallend”, wat betekent dat volgens jullie? Wat betekent dat voor jullie gebruik van de grafiek? Hoe zou je een onopvallend profiel volgens jullie weer moeten geven (eventueel niet weergeven)?

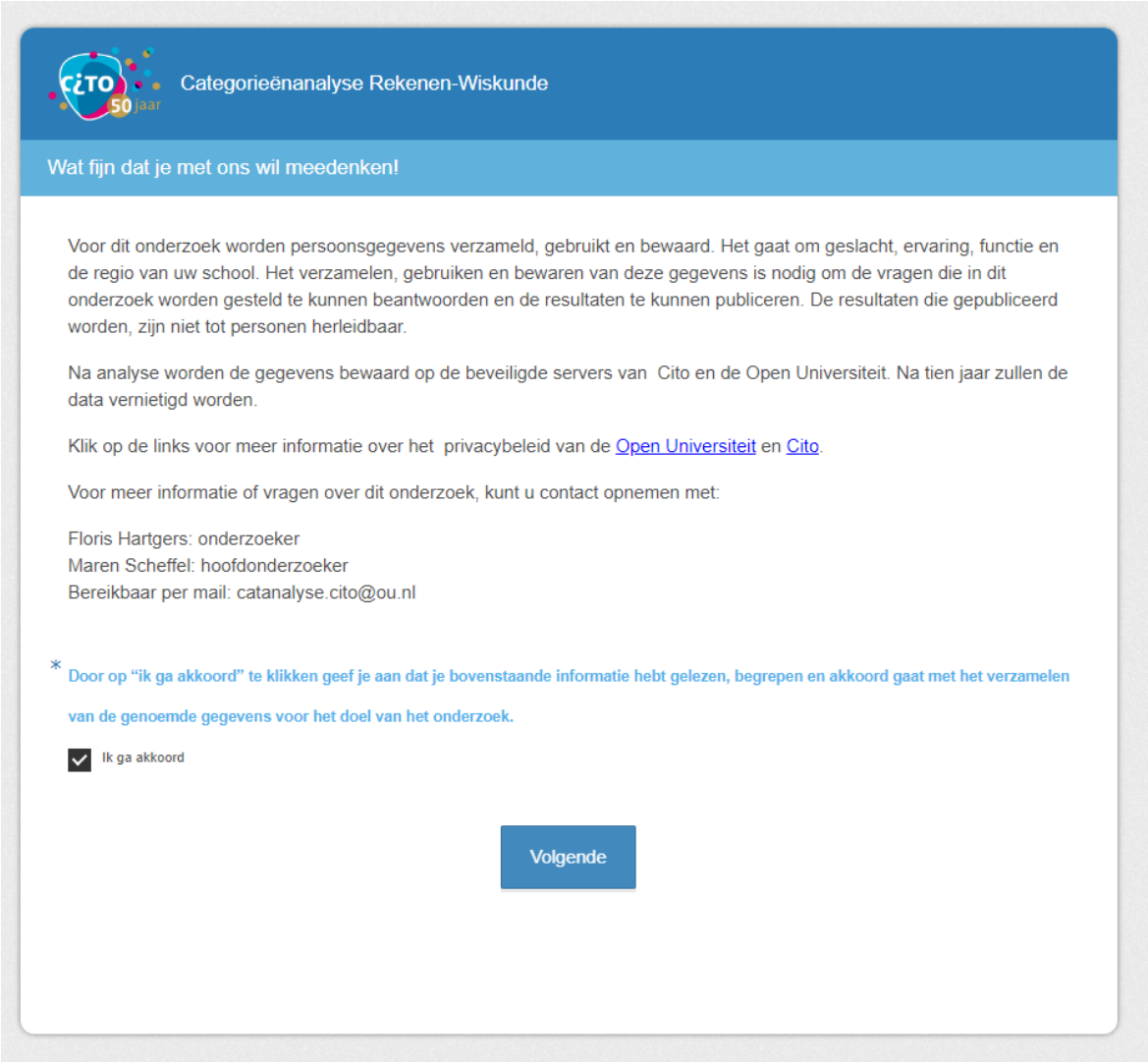
In de alternatieve weergaven is minder informatie zichtbaar, wat vind je hiervan? Welke informatie is zinvol volgens jullie en welke zou weggelaten kunnen worden?

Veranderde elementen:

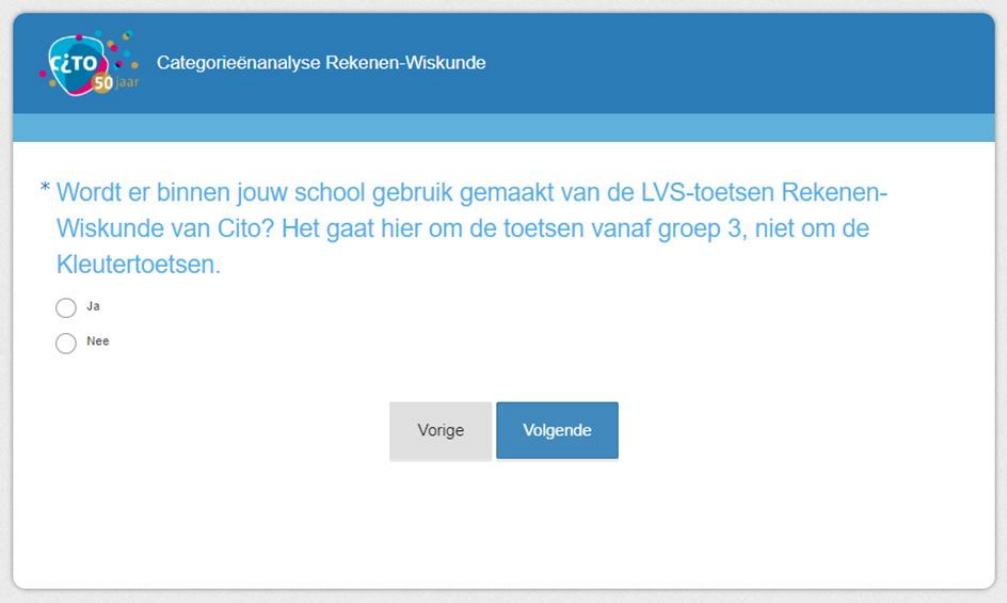
- Aanduiding niet opvallend, opvallend, zeer opvallend, naar boven verplaatst, groter, zonder grafiek.
- Staven in staafgrafiek enkel variërend in hoogte, niet meer in oppervlakte
- Geen rode of zwarte cijfers
- Kleuren aangepast
- Weglaten getallen bij de balk
- Weglaten tabel
- Afkortingen onder grafiek vervangen door gehele term

Zijn er nog andere verbeteringen/veranderingen mogelijk die je beter informatie zouden geven?

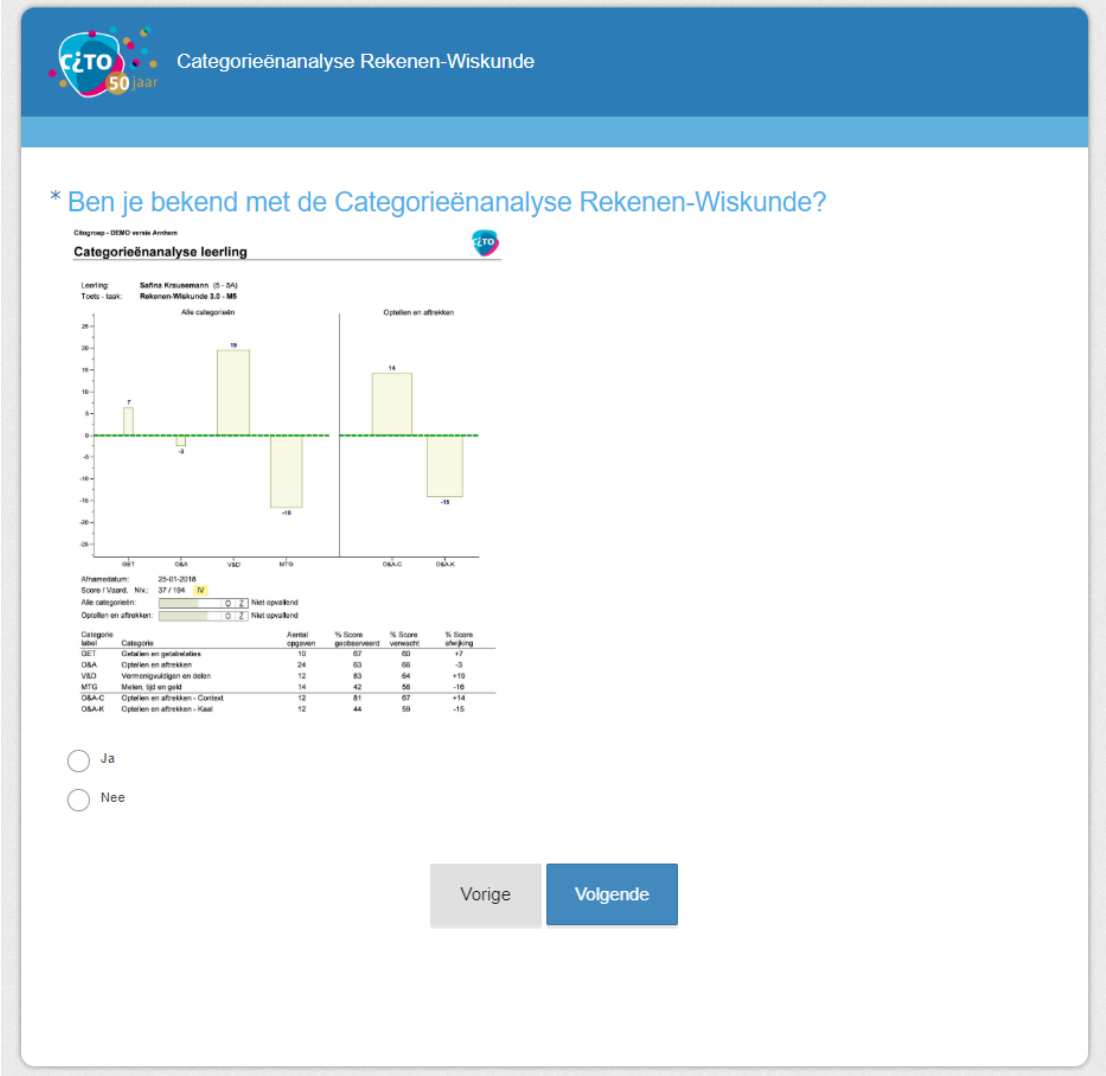
Appendix C – Survey



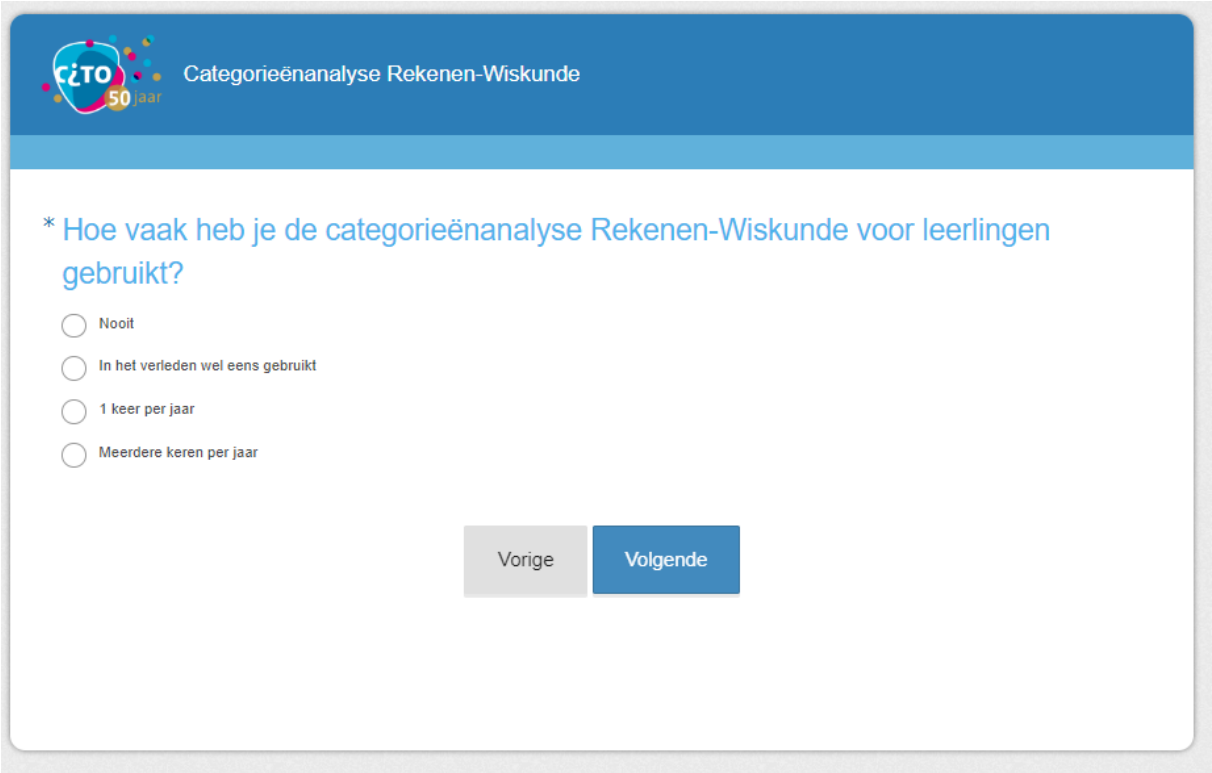
Appendix C - informed consent message



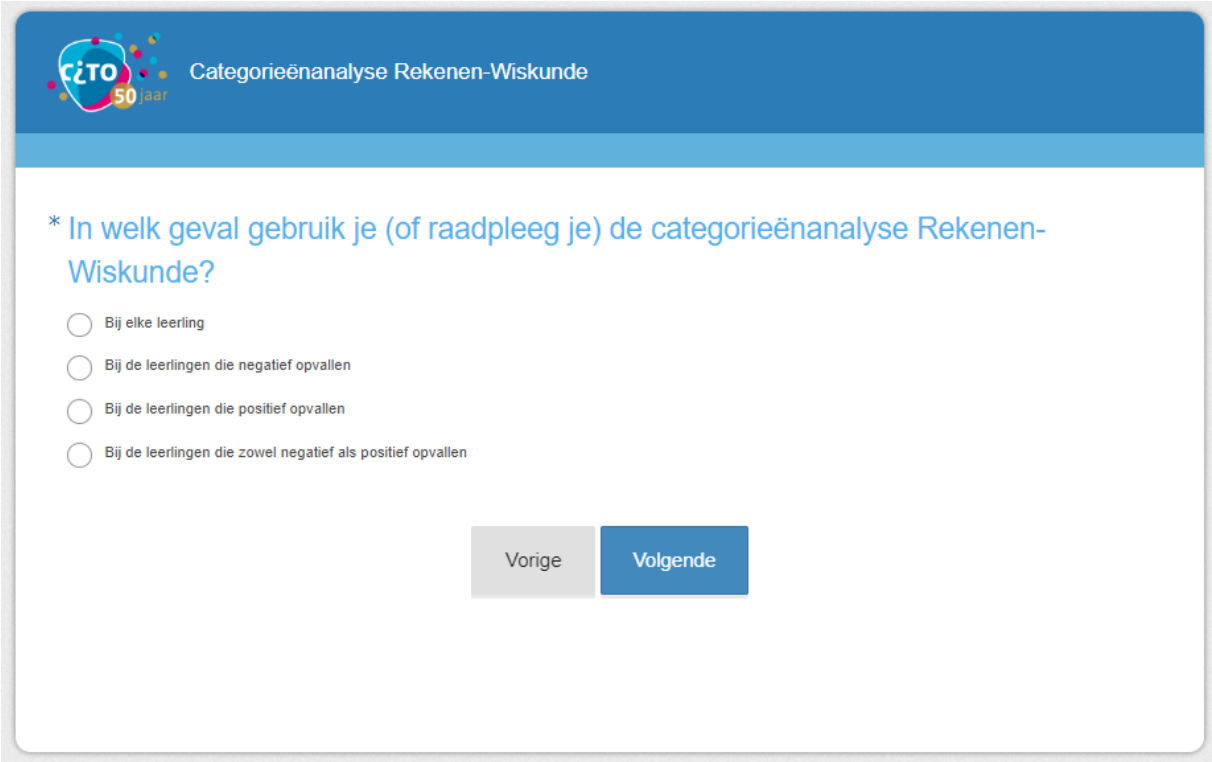
Appendix C - Question 1



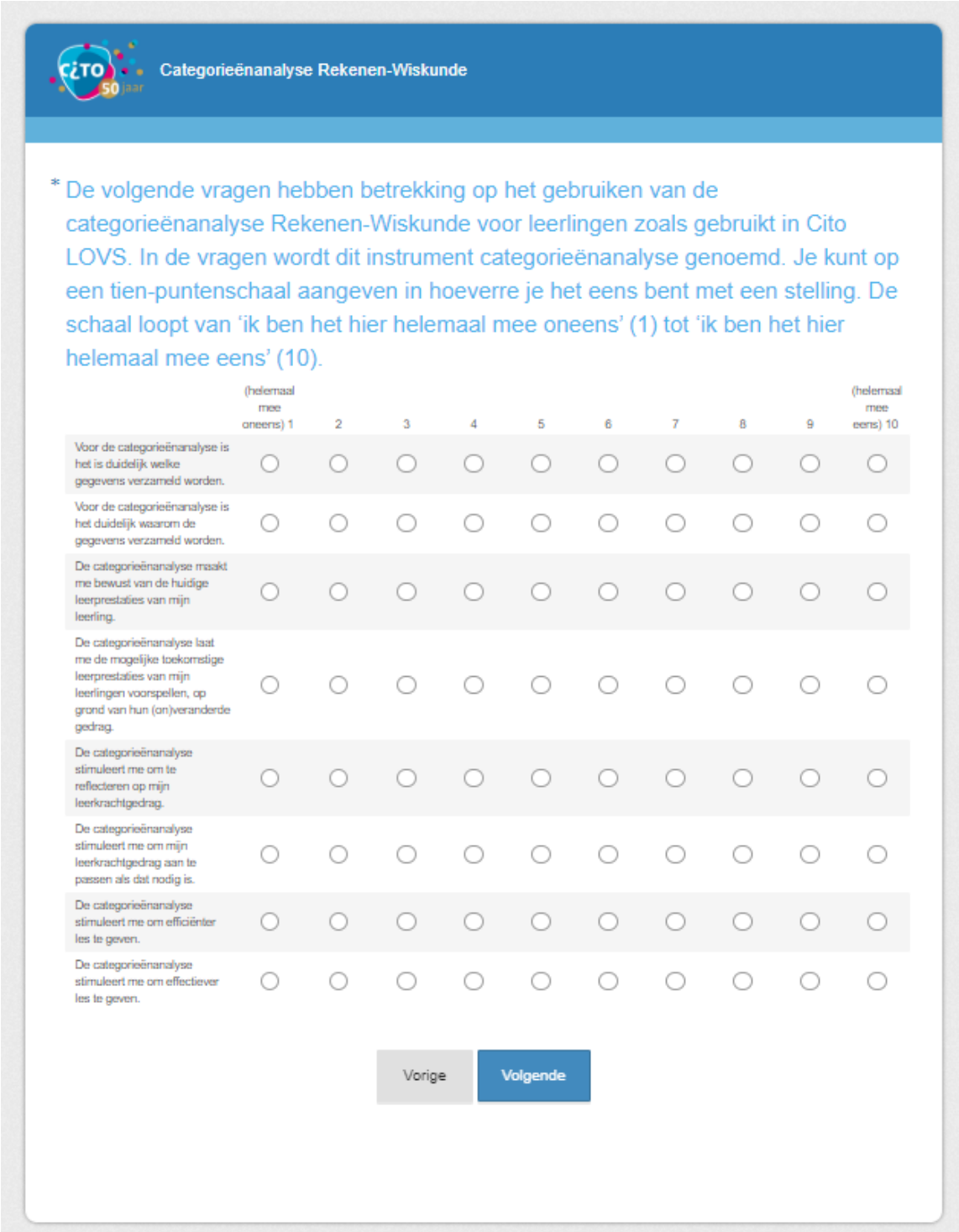
Appendix C - Question 2



Appendix C - Question 3



Appendix C - Question 4



Categorieënanalyse Rekenen-Wiskunde

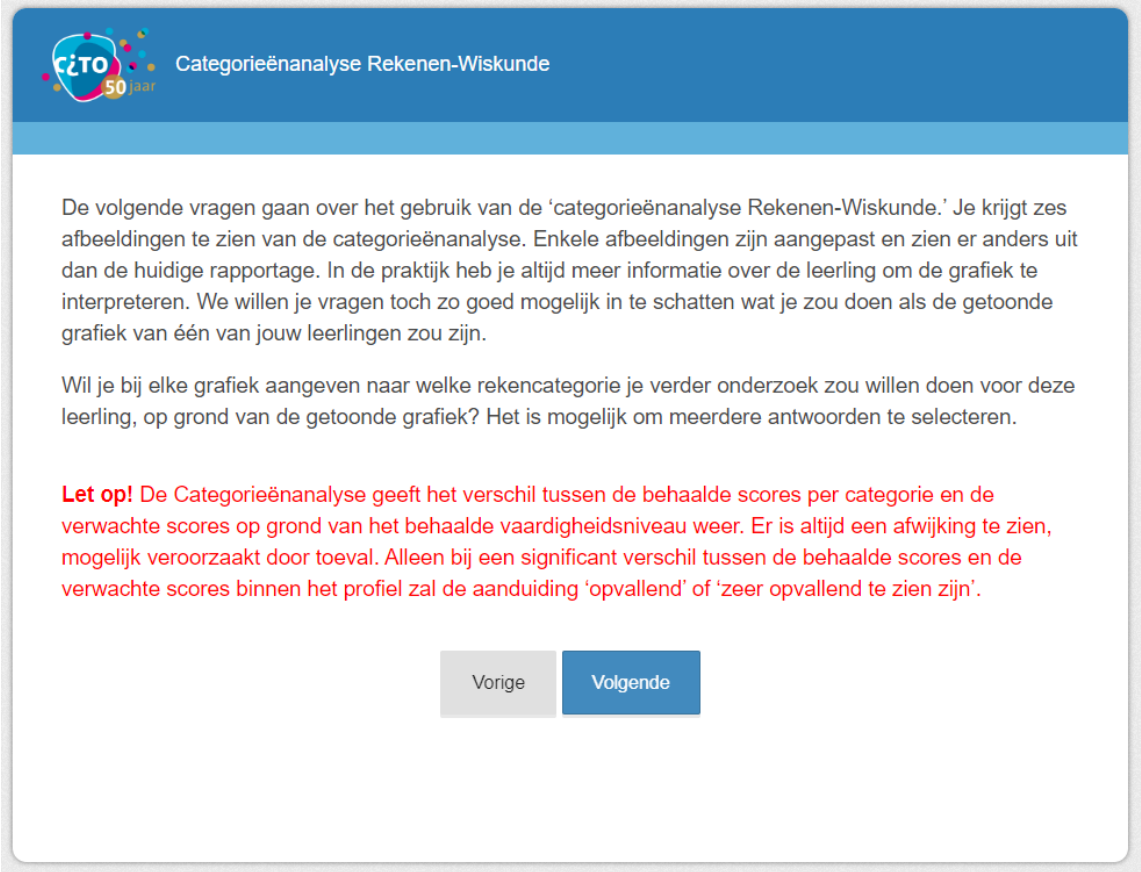
* De volgende vragen hebben betrekking op het gebruiken van de categorieënanalyse Rekenen-Wiskunde voor leerlingen zoals gebruikt in Cito LOVS. In de vragen wordt dit instrument categorieënanalyse genoemd. Je kunt op een tien-puntenschaal aangeven in hoeverre je het eens bent met een stelling. De schaal loopt van 'ik ben het hier helemaal mee oneens' (1) tot 'ik ben het hier helemaal mee eens' (10).

	(helemaal mee oneens) 1	2	3	4	5	6	7	8	9	(helemaal mee eens) 10
Voor de categorieënanalyse is het is duidelijk welke gegevens verzameld worden.	○	○	○	○	○	○	○	○	○	○
Voor de categorieënanalyse is het duidelijk waarom de gegevens verzameld worden.	○	○	○	○	○	○	○	○	○	○
De categorieënanalyse maakt me bewust van de huidige leerprestaties van mijn leerling.	○	○	○	○	○	○	○	○	○	○
De categorieënanalyse laat me de mogelijke toekomstige leerprestaties van mijn leerlingen voorspellen, op grond van hun (on)veranderde gedrag.	○	○	○	○	○	○	○	○	○	○
De categorieënanalyse stimuleert me om te reflecteren op mijn leerkrachtgedrag.	○	○	○	○	○	○	○	○	○	○
De categorieënanalyse stimuleert me om mijn leerkrachtgedrag aan te passen als dat nodig is.	○	○	○	○	○	○	○	○	○	○
De categorieënanalyse stimuleert me om efficiënter les te geven.	○	○	○	○	○	○	○	○	○	○
De categorieënanalyse stimuleert me om effectiever les te geven.	○	○	○	○	○	○	○	○	○	○

Vorige Volgende

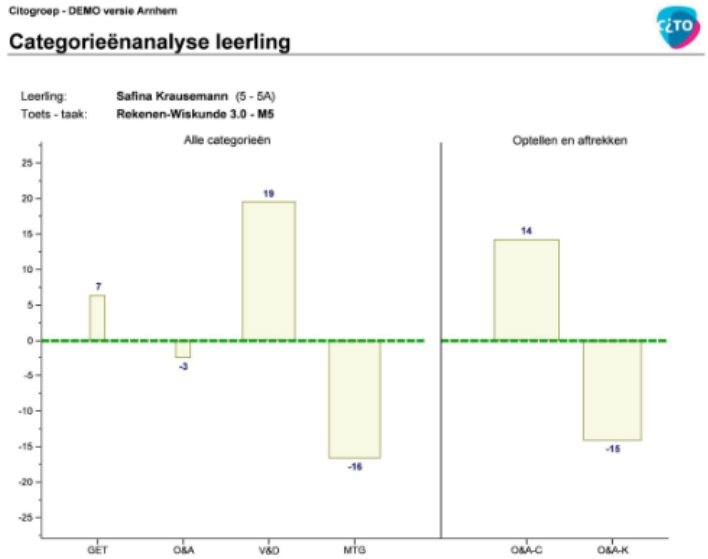
Appendix C - Question 5: translated EFLA questionnaire

Appendix C - instruction, without disclaimer



Appendix C - instruction, with disclaimer

* Ik zou bij deze grafieken één of meer van de volgende categorieën verder willen onderzoeken:



Citogroep - DEMO versie Arnhem

Categorieënanalyse leerling

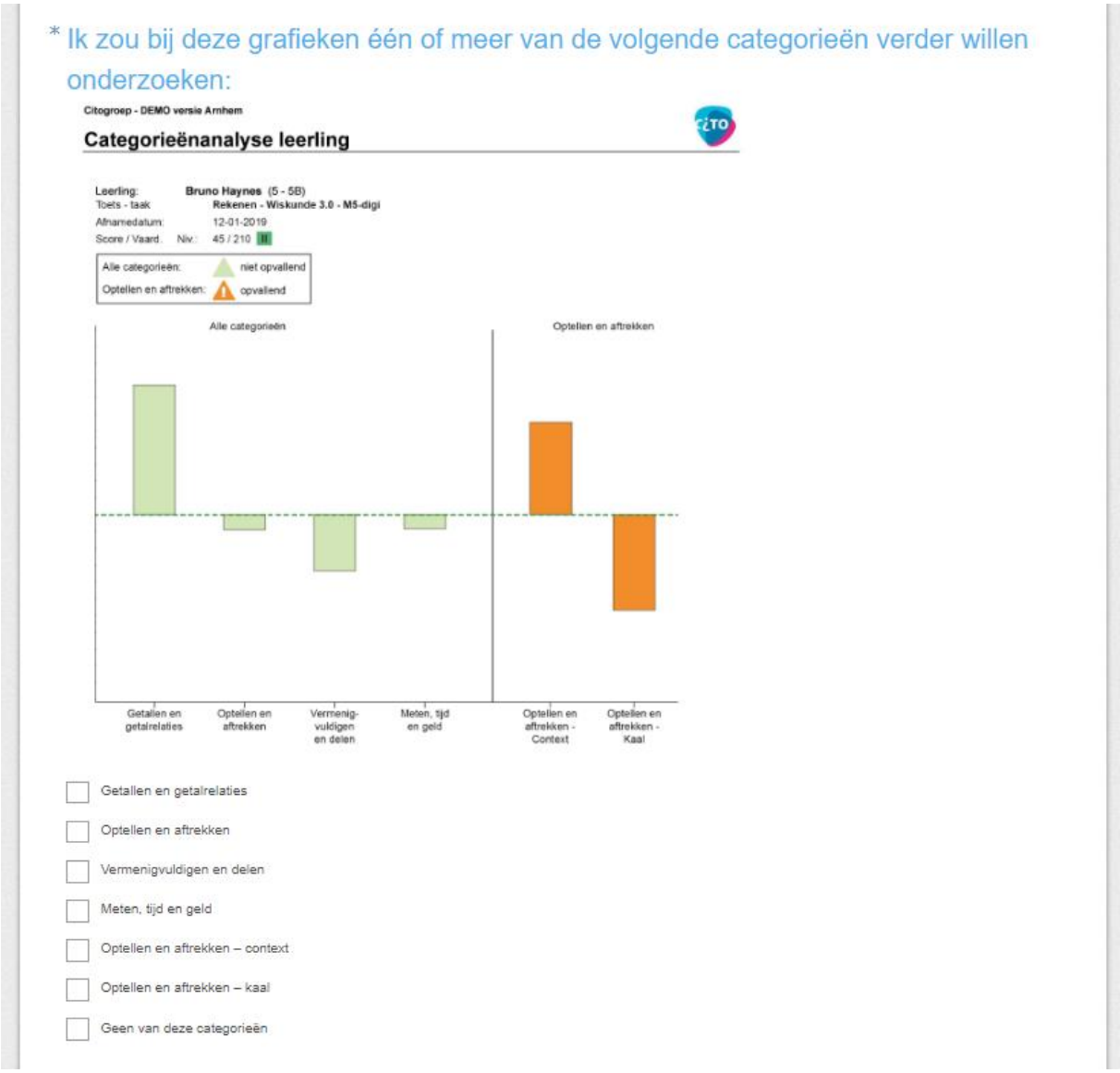
Leerling: **Safina Krausemann** (5 - 5A)
Toets - taak: **Rekenen-Wiskunde 3.0 - M5**

Afnamedatum: 25-01-2018
Score / Vaard. Niv.: 37 / 194 IV
Alle categorieën: O Z Niet opvallend
Optellen en aftrekken: O Z Niet opvallend

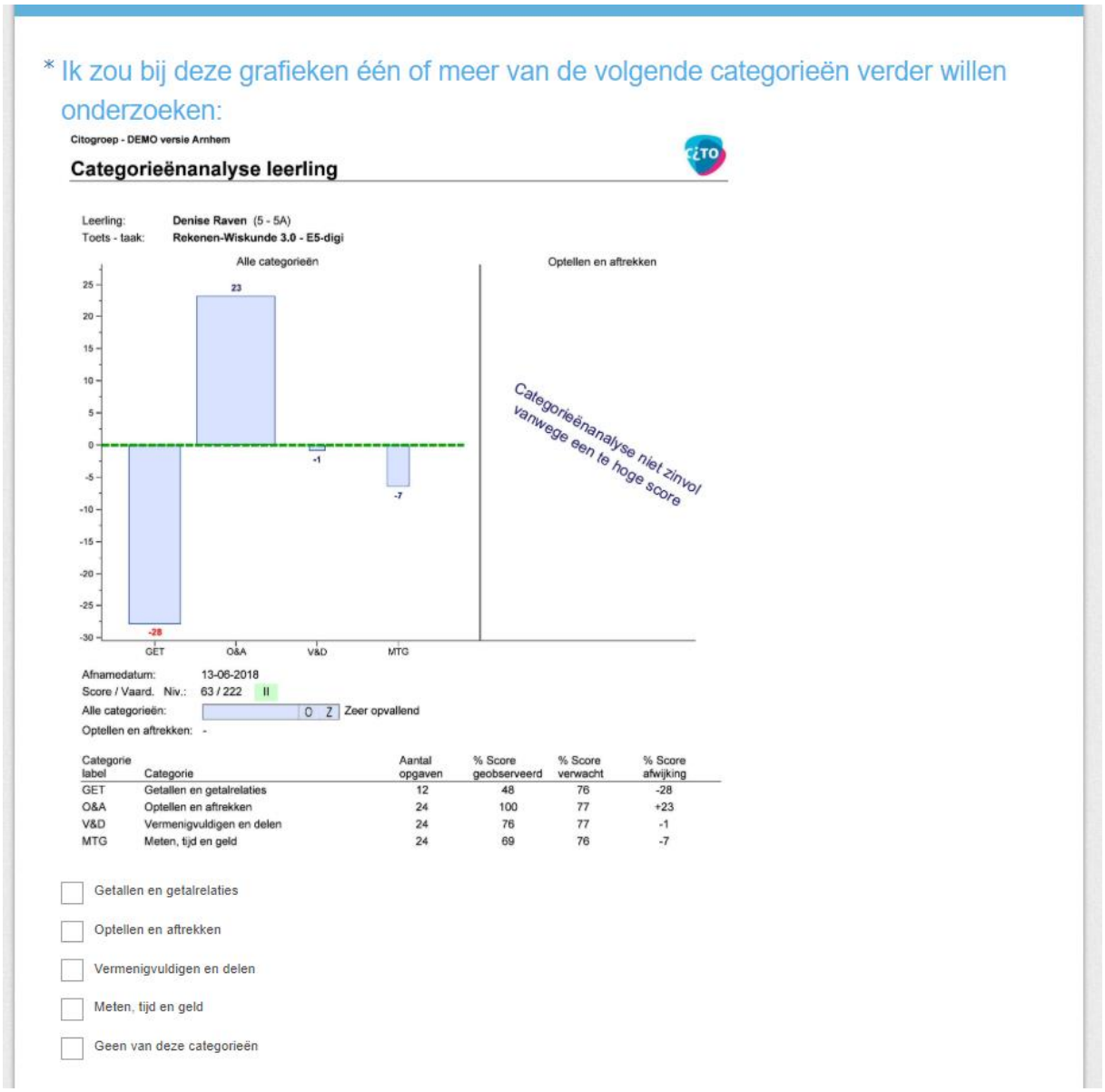
Categorie label	Categorie	Aantal opgaven	% Score geobserveerd	% Score verwacht	% Score afwijking
GET	Getallen en getalrelaties	10	67	60	+7
O&A	Optellen en aftrekken	24	63	66	-3
V&D	Vermenigvuldigen en delen	12	83	64	+19
MTG	Meten, tijd en geld	14	42	58	-16
O&A-C	Optellen en aftrekken - Context	12	81	67	+14
O&A-K	Optellen en aftrekken - Kaal	12	44	59	-15

- [ ] Getallen en getalrelaties
- [ ] Optellen en aftrekken
- [ ] Vermenigvuldigen en delen
- [ ] Meten, tijd en geld
- [ ] Optellen en aftrekken – context
- [ ] Optellen en aftrekken – kaal
- [ ] Geen van deze categorieën

Appendix C - Question 6: Dashboard 1

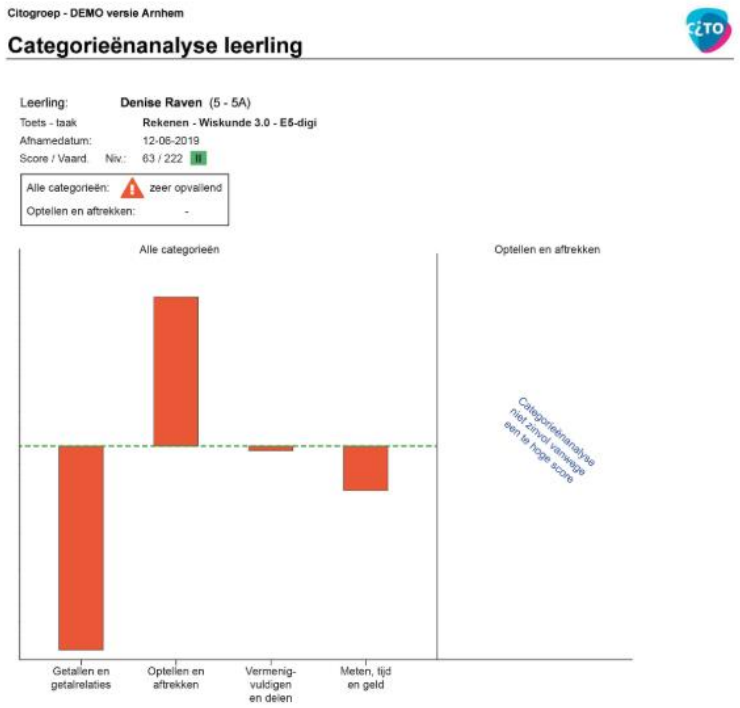


Appendix C – Question 6: Dashboard 1A



Appendix C – Question 6: Dashboard 2

* Ik zou bij deze grafieken één of meer van de volgende categorieën verder willen onderzoeken:



- [ ] Getallen en getalrelaties
- [ ] Optellen en aftrekken
- [ ] Vermenigvuldigen en delen
- [ ] Meten, tijd en geld
- [ ] Geen van deze categorieën

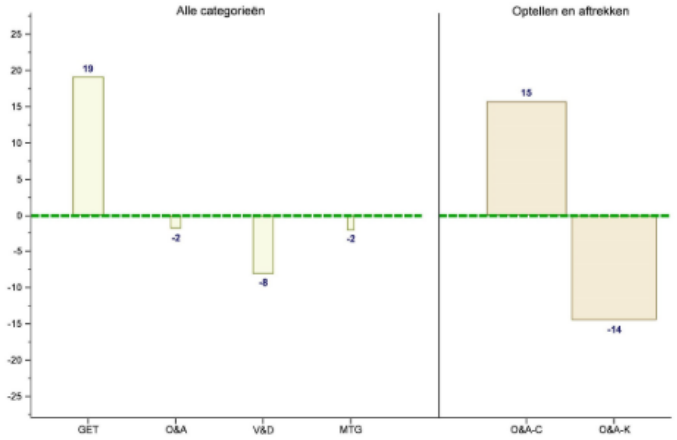
Appendix C – Question6: Dashboard 2A

* Ik zou bij deze grafieken één of meer van de volgende categorieën verder willen onderzoeken:

Citogroep - DEMO versie Arnhem

Categorieënanalyse leerling

Leerling: **Bruno Haynes** (5 - 5B)
Toets - taak: **Rekenen-Wiskunde 3.0 - M5-digi**



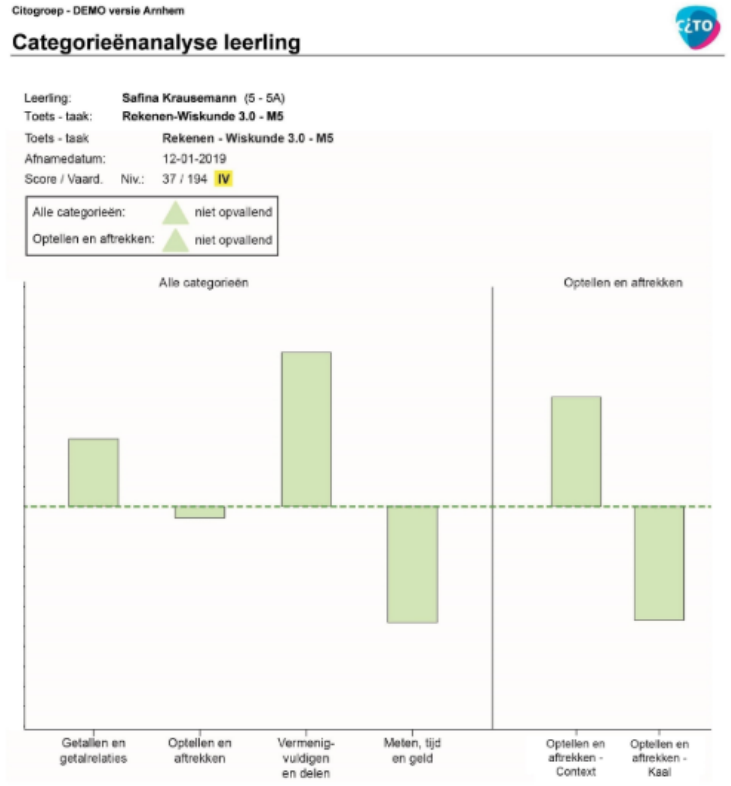
Afnamedatum: 23-01-2018
Score / Vaard. Niv.: 45 / 210 II
Alle categorieën: O Z Niet opvallend
Optellen en aftrekken: O Z Opvallend

Categorie label	Categorie	Aantal opgaven	% Score geobserveerd	% Score verwacht	% Score afwijking
GET	Getallen en getalrelaties	10	88	69	+19
O&A	Optellen en aftrekken	24	76	78	-2
V&D	Vermenigvuldigen en delen	12	67	75	-8
MTG	Meten, tijd en geld	14	75	77	-2
O&A-C	Optellen en aftrekken - Context	12	90	75	+15
O&A-K	Optellen en aftrekken - Kaal	12	63	77	-14

- [ ] Getallen en getalrelaties
- [ ] Optellen en aftrekken
- [ ] Vermenigvuldigen en delen
- [ ] Meten, tijd en geld
- [ ] Optellen en aftrekken – context
- [ ] Optellen en aftrekken – kaal
- [ ] Geen van deze categorieën

Appendix C - Question 6: Dashboard 3

* Ik zou bij deze grafieken één of meer van de volgende categorieën verder willen onderzoeken:



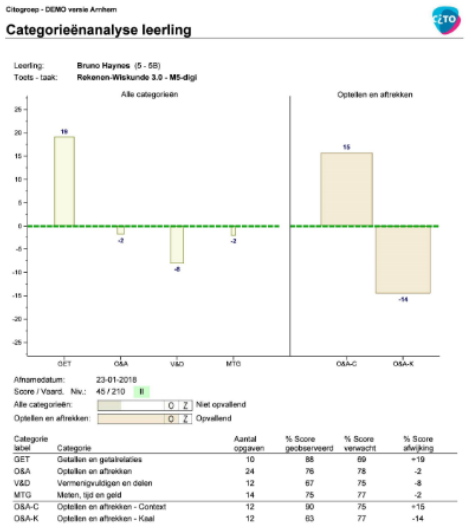
- [ ] Getallen en getalrelaties
- [ ] Optellen en aftrekken
- [ ] Vermenigvuldigen en delen
- [ ] Meten, tijd en geld
- [ ] Optellen en aftrekken – context
- [ ] Optellen en aftrekken – kaal
- [ ] Geen van deze categorieën

Appendix C - Question 6: Dashboard 3A

Categorieënanalyse Rekenen-Wiskunde

De volgende vragen gaan over elementen uit de categorieënanalyse.

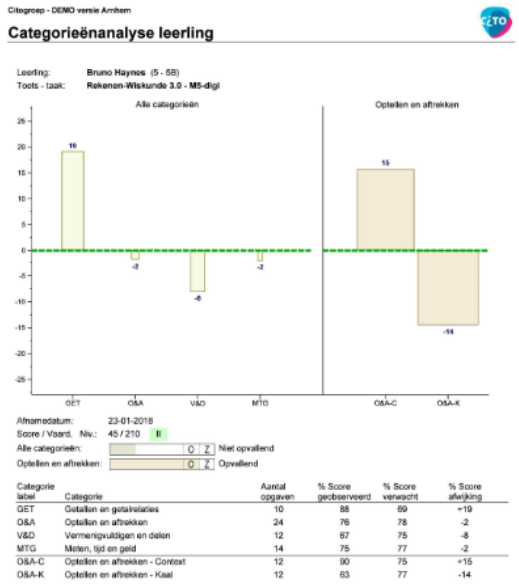
Wat denk jij dat betekenis van de groene lijn is? Je kunt één antwoord kiezen.



- ○ De groene stippellijn geeft de norm aan waaraan de leerling moet voldoen om een categorie voldoende te maken
- ○ De groene stippellijn geeft het verwachte profiel weer op grond van de vaardigheidsgroei ten opzichte van de vorige toets
- ○ De groene stippellijn geeft het verwachte profiel weer op grond van de behaalde vaardigheidsscore
- ○ De groene stippellijn geeft de norm weer waaraan leerlingen met een bepaalde Cito score (I tm. V) moeten voldoen

Appendix C - Question 7: knowledge 1

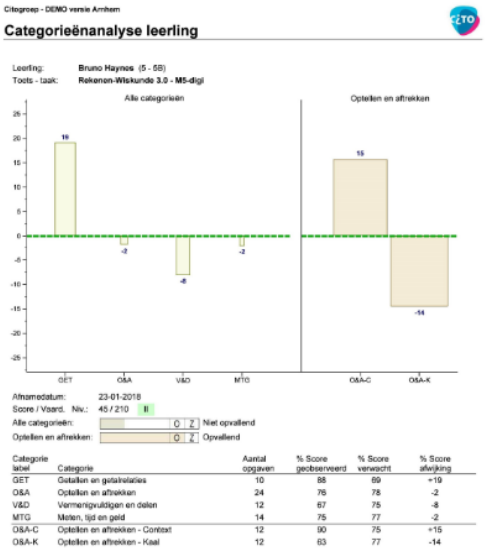
Wat denk jij dat de hoogte van de staven aangeeft? Je kunt één antwoord kiezen.



- ◯ De hoogte van de staven geeft het verschil aan tussen de behaalde score op die categorie en de verwachte score van leerlingen met ongeveer dezelfde vaardigheidsscore
- ◯ De hoogte van de staven geeft de score weer die de leerling heeft behaald op een bepaalde categorie
- ◯ De hoogte van de staven geeft aan in welke mate een leerling een voldoende of onvoldoende score heeft behaald op een bepaalde categorie
- ◯ De hoogte van de staven geeft de vaardigheidsgroei aan die de leerling heeft behaald ten opzichte van de vorige toets
- ◯ Ik weet niet wat de hoogte van de staven betekent

Appendix C - Question 7: knowledge 2

Wat denk jij dat de oppervlakte van de staven aangeeft? Je kunt één antwoord kiezen.

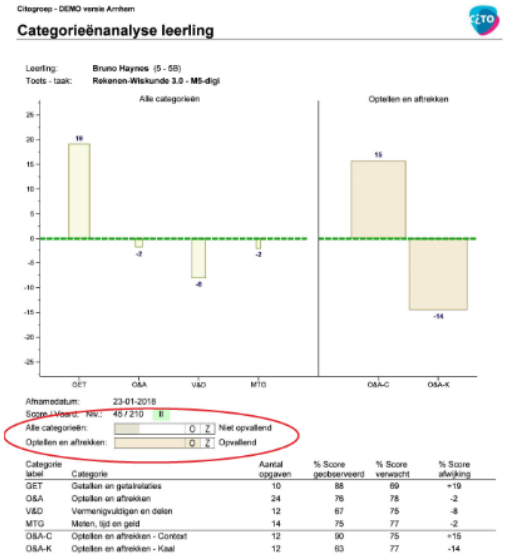


Categorie label	Categorie	Aantal opgaven	% Score geobserveerd	% Score verwacht	% Score afwijking
GET	Getallen en getalrelaties	10	88	69	+19
O&A	Optellen en aftrekken	24	76	78	-2
V&D	Vermenigvuldigen en delen	12	67	75	-8
MTG	Meten, tijd en geld	14	75	77	-2
O&A-C	Optellen en aftrekken - Context	12	90	75	+15
O&A-K	Optellen en aftrekken - Kaal	12	63	77	-14

- ◯ De oppervlakte van de staven geeft aan hoe belangrijk een categorie is voor leerlingen in dit leerjaar
- ◯ De oppervlakte van de staven geeft aan hoe zwaar de staven onderling wegen binnen het profiel
- ◯ De oppervlakte van de staven geeft aan hoeveel vragen goed of fout gemaakt zijn
- ◯ De oppervlakte van de staven geeft aan hoeveel een leerling gegroeid is in vaardigheidsniveau ten opzichte van de vorige toets
- ◯ Ik weet niet wat de oppervlakte van de staven betekent

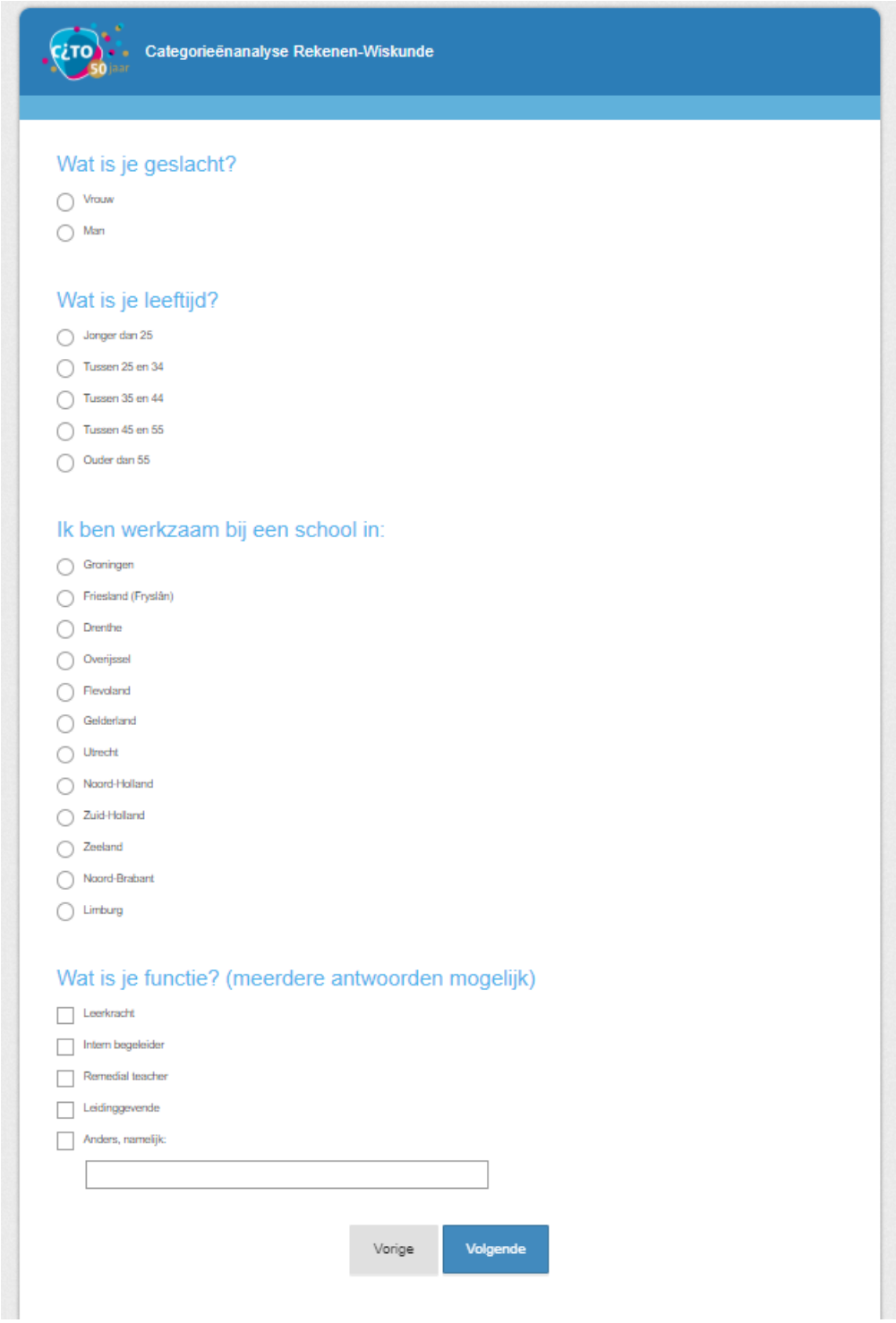
Appendix C - Question 7: knowledge 3

Wat denk jij dat de betekenis van de grafiek signaal is? De grafiek signaal is omcirkeld in de afbeelding. Je kunt één antwoord kiezen.



- Het signaal 'opvallend' of 'zeer opvallend' wordt gegeven als leerlingen onvoldoende groei hebben laten zien in vaardigheidsscore
- Het signaal 'opvallend' of 'zeer opvallend' wordt gegeven als een leerling zorgwekkend laag scoort op één of categorieën
- Het signaal 'opvallend' of 'zeer opvallend' wordt gegeven als deze categorie in de groep opvallend zwak wordt gemaakt
- Het signaal 'opvallend' of 'zeer opvallend' wordt gegeven als het verschil tussen het geobserveerde profiel en het verwachte profiel statistisch significant is
- Ik weet niet wat het signaal 'opvallend' of 'zeer opvallend' betekent

Appendix C - Question 7: knowledge 4



Appendix C - Question 8: background questions

Appendix C - end message